On consistency and sparsity for sliced inverse regression in high dimensions

Submitted to the Annals of Statistics
arXiv: arXiv:0000.0000

ON CONSISTENCY AND SPARSITY FOR SLICED INVERSE REGRESSION IN HIGH DIMENSIONS

By Qian Lin[§*] Zhigen Zhao[¶*] and Jun S. Liu[§*]

Harvard University[§]
Temple University[¶]

We provide here a framework to analyze the phase transition phenomenon of slice inverse regression (SIR), a supervised dimension reduction technique introduced by Li [1991]. Under mild conditions, the asymptotic ratio $\rho = \lim p/n$ is the phase transition parameter and the SIR estimator is consistent if and only if $\rho = 0$. When dimension p is greater than n, we propose a diagonal thresholding screening SIR (DT-SIR) algorithm. This method provides us with an estimate of the eigen-space of the covariance matrix of the conditional expectation $var(\mathbb{E}[\boldsymbol{x}|y])$. The desired dimension reduction space is then obtained by multiplying the inverse of the covariance matrix on the eigen-space. Under certain sparsity assumptions on both the covariance matrix of predictors and the loadings of the directions, we prove the consistency of DT-SIR in estimating the dimension reduction space in high dimensional data analysis. Extensive numerical experiments demonstrate superior performances of the proposed method in comparison to its competitors.

1. Introduction. For a continuous multivariate random variable $(y, \boldsymbol{x})$ where $\boldsymbol{x} \in \mathbb{R}^p$ and $y \in \mathbb{R}$, a subspace $\mathcal{S}' \subset \mathbb{R}^p$ is called the effective dimension reduction (EDR) space if $y \perp\!\!\!\perp \boldsymbol{x}|P_{\mathcal{S}'}(\boldsymbol{x})$ where $\perp\!\!\!\perp$ stands for independence. Under mild conditions (Cook [1996]), the intersection of all the EDR spaces is again an EDR space, which is denoted as $\mathcal{S}$ and called the central space. Many algorithms were proposed to find such subspace $\mathcal{S}$ under the assumption $d = dim\mathcal{S} \ll p$. This line of research is commonly known as sufficient dimension reduction. The Sliced Inverse Regression (SIR, Li [1991]) is the first, yet the most widely used method in sufficient dimension reduction, due to its simplicity, computational efficiency and generality. The asymptotic properties of SIR are of particular interest in the last two decades. The consistency of SIR has been proved for fixed p in Li [1991], Hsing and Carroll

*Lin's research is supported by the Center of Mathematical Sciences and Applications at Harvard University. Zhao's research is supported by the NSF Grant DMS-1208735. Liu's research is supported by the NSF Grant DMS-1120368 and NIH Grant R01 GM113242-01

MSC 2010 subject classifications: Primary 62J02; secondary 62H25

Keywords and phrases: dimension reduction, random matrix theory, sliced inverse regression

[1992], Zhu and Ng [1995] and Zhu and Fang [1996]. Later, Zhu et al. [2006] have obtained the consistency if $p = o(\sqrt{n})$. A similar restriction also appears in two recent work (see Zhong et al. [2012] and Jiang and Liu [2014]). When $p > n$, a common strategy pursued by many recent researchers is to make sparsity assumptions that only a few predictors play a role in explaining and predicting y and apply various regularization methods. For instance, Li and Nachtsheim [2006], Li [2007] and Yu et al. [2013] applied LASSO (Tibshirani [1996]), Dantzig selector (Candes and Tao [2007]) and elastic net (Zou and Hastie [2005]) respectively to solve the generalized eigenvalue problems raised by a variety of SDR algorithms.

However, a piece of jigsaw is missing in the understanding of SIR. If the dimension p diverges as n increases, when will the SIR break down? A similar question has been asked for a variety of SDR estimates in Cook et al. [2012]. In this paper, we prove that, under certain technical assumptions, the SIR estimator is consistent if and only if $\rho = \lim \frac{p}{n} = 0$. Such a result on inconsistency provides theoretical justifications for imposing certain structural assumption, such as sparsity, in high dimensional settings. This behavior of SIR in high dimension, which will be called the phase transition phenomenon, is similar to that of the principal component analysis (PCA), an unsupervised counterpart of SIR. This extension is, however, by no means trivial. After all the samples $(y_i, \boldsymbol{x}_i)$ are sliced into H bins according to the order statistics of y_i , the sliced samples are neither independent nor identically distributed. This difference increases the difficulty significantly. In this paper, we provide a new framework to study the phase transition behaviour of SIR. The technical tools developed here can potentially be extended to study the phase transition behaviour of other SDR estimators.

The second part of the article aims at extending the original SIR to the scenario with ultra-high dimension ($p = o(\exp(n^{\xi}))$). Based on equation (3) in Section 2, the central space can be estimated by the column space of $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} col(\widehat{\boldsymbol{V}}_H)$, where $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}$ is any consistent estimate of the precision matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$ and $col(\widehat{\boldsymbol{V}}_H)$ is the estimate of the space $col(var(\mathbb{E}[\boldsymbol{x}|y]))$. To estimate the column space of $var\,(\mathbb{E}[\boldsymbol{x}|y])$, we propose a diagonal screening procedure based on new univariate statistics $var_H(\boldsymbol{x}(k))$, which are the diagonal elements of $var\,(\mathbb{E}[\boldsymbol{x}|y])$, motivated by recent work in sparse PCA (Johnstone and Lu [2004]). After ranking the predictors according to the magnitude of $var_H(\boldsymbol{x}(k))$ decreasingly, we choose the set $\mathcal{I}$ consisting of the first R predictors as active predictors. The SIR procedure is subsequently applied to these selected predictors to estimate the d-dimensional column space of $var(\mathbb{E}[\boldsymbol{x}^{\mathcal{I}}|y])$ by $col(\widehat{\boldsymbol{V}}_H^{\mathcal{I}})$ where $\widehat{\boldsymbol{V}}_H^{\mathcal{I}}$ is the matrix formed by the top d eigenvectors of $\widehat{\boldsymbol{\Lambda}}_H^{\mathcal{I}}$. We embed $\widehat{\boldsymbol{V}}_H^{\mathcal{I}}$ into $\mathbb{R}^{p\times d}$ by filling in 0's for entries outside

the chosen row set $\mathcal{I}$, and denote this new matrix by $e(\widehat{\boldsymbol{V}}_H^{\mathcal{I}})$. The estimate of the central space is defined to be $col(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} e(\widehat{\boldsymbol{V}}_H^{\mathcal{I}}))$. We name this two-stage algorithm as **D**iagonal **T**hresholding **SIR** (DT-SIR), and prove that DT-SIR is consistent in estimating the central space under certain regularity conditions. Extensive simulation studies show that DT-SIR performs better than its competitors and is computationally efficient.

The rest of the paper is organized as follows. In Section 2, we briefly describe the SIR procedure and introduce the notations. In Section 3, after a brief review of existing asymptotic results of SIR procedure, we state Theorems 2 and 3 to discuss the phase transition phenomenon of SIR. In Section 4, we propose the DT-SIR method and show that DT-SIR is consistent in high dimensional data analysis. In Section 5, we provide simulation studies to compare DT-SIR with its competitors. Concluding remarks and discussions are put in Section 6. All the proofs are presented in appendices.

2. Preliminaries and notations.

2.1. *Sliced inverse regression* Consider the multiple index model

$$y = f(\boldsymbol{\beta}_1^\tau \boldsymbol{x}, \cdots, \boldsymbol{\beta}_d^\tau \boldsymbol{x}, \epsilon) \tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^p$, ϵ is the noise and f is an unknown link function. Without loss of generality, we assume that $\mathbb{E}[\boldsymbol{x}] = 0 \in \mathbb{R}^p$. Although the $p \times d$ matrix $\boldsymbol{V} = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_d)$ is not identifiable, the space spanned by the $\boldsymbol{\beta}$'s, which is called the column space of $\boldsymbol{V}$ and denoted by $col(\boldsymbol{V})$, might be identified. Li [1991] proposed the *Sliced Inverse Regression* (SIR) procedure to estimate the central space $col(\boldsymbol{V})$ without knowing $f(\cdot)$, which can be briefly summarized as follows: Given n i.i.d. samples $(y_i, \boldsymbol{x}_i)$, $i = 1, \cdots, n$, SIR first divides them into H equal-sized slices according to the order statistics $y_{(i)}$.[1] We re-express the data as $y_{h,j}$ and $\boldsymbol{x}_{h,j}$, where (h, j) is the double subscript in which h refers to the slice number and j refers to the order number of a sample in the h-th slice, i.e.,

$$y_{h,j} = y_{(c(h-1)+j)}, \qquad \boldsymbol{x}_{h,j} = \boldsymbol{x}_{(c(h-1)+j)}.$$

Here $\boldsymbol{x}_{(k)}$ is the concomitant of $y_{(k)}$. Let the sample mean in the h-th slice be $\overline{\boldsymbol{x}}_{h,\cdot}$, and let the mean of all the samples be $\overline{\overline{\boldsymbol{x}}}$. Then, $\boldsymbol{\Lambda}_p \triangleq var(\mathbb{E}[\boldsymbol{x}|y])$ can be estimated by:

$$\widehat{\boldsymbol{\Lambda}}_H = \frac{1}{H} \sum_{h=1}^{H} \bar{\boldsymbol{x}}_{h,\cdot} \bar{\boldsymbol{x}}_{h,\cdot}^\tau. \tag{2}$$

[1] To ease notations and arguments, we assume that $n = cH$ and $H = o\left(\log(n) \wedge \log(p)\right)$ throughout the article.

Based on the observation that

$$col(\mathbf{\Lambda}) = \mathbf{\Sigma}_x col(\boldsymbol{V}), \tag{3}$$

the SIR then estimates the central space $col(\boldsymbol{V})$ by $\widehat{\Sigma}_{\boldsymbol{x}}^{-1} col(\widehat{\boldsymbol{V}}_H)$ where $\widehat{V}_H$ is the matrix formed by the top d eigenvectors of $\hat{\mathbf{\Lambda}}_H$.. Throughout the article, we assume that d is fixed and the d-th largest eigenvalue λ_d of $\mathbf{\Lambda}_p$ is bounded away from 0 when $n, p \to \infty$. In order for SIR to result in a consistent estimate of the central space, Li [1991] imposed the the following two conditions:

- (**A1**). **Linearity condition:** For any $\boldsymbol{\xi} \in \mathbb{R}^p$, $\mathbb{E}[\boldsymbol{\xi}^\tau \boldsymbol{x} | \boldsymbol{\beta}_1^\tau \boldsymbol{x}, \cdots, \boldsymbol{\beta}_d^\tau \boldsymbol{x}]$ is a linear combination of $\boldsymbol{\beta}_1^\tau \boldsymbol{x}, \cdots, \boldsymbol{\beta}_d^\tau \boldsymbol{x}$.
- (**A2**). **Coverage condition:** The dimension of the space spanned by the central curve equals the dimension of the central space, i.e., $d' = d$.

2.2. *Further Notations.* Let S_h be the h-th interval $(y_{h-1,c}, y_{h,c}]$ for $2 \le h \le H-1$, $S_1 = (-\infty, y_{1,c}]$ and $S_H = (y_{H-1,c}, \infty)$. Note that these intervals depend on the order statistics $y_{(i)}$ and are thus random. For any ω in the product sample space, we define a random variable $\delta_h = \delta_h(\omega) = \int_{y \in S_h(\omega)} f(y) dy$ where $f(y)$ is the density function of y. For $\mathcal{I} \subset \{1, \cdots, n\}, \mathcal{J} \subset \{1, \cdots, p\}$ and a $n \times p$ matrix $\boldsymbol{A}$, $\boldsymbol{A}^{\mathcal{I},\mathcal{J}}$ denotes the $|\mathcal{I}| \times |\mathcal{J}|$ sub-matrix formed by restricting the rows of $\boldsymbol{A}$ to $\mathcal{I}$ and columns to $\mathcal{J}$. In articular, $\boldsymbol{A}^{-,\mathcal{J}}$ denotes the sub-matrix formed by restricting the columns to $\mathcal{J}$; For a matrix $\boldsymbol{B} = \boldsymbol{A}^{\mathcal{I},\mathcal{J}} \in \mathbb{R}^{|\mathcal{I}| \times |\mathcal{J}|}$, we embed it into $\mathbb{R}^{p \times p}$ by putting 0 on entries outside $\mathcal{I} \times \mathcal{J}$ and denote the new matrix as $e(\boldsymbol{B})$. Similar notations apply to vectors. For two positive numbers a and b, we let $a \vee b \equiv \max\{a, b\}$ and let $a \wedge b \equiv \min\{a, b\}$. Let $\tau(x, t) = x \times 1(|x| > t)$ be the hard thresholding function. Throughout the article, C, C_1 and C_2 are used to denote generic absolute constants, though the actual value may vary from case to case. For a vector $\boldsymbol{x}$, we denote its k-th entry as $\boldsymbol{x}(k)$. Let $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ be two vectors with the same dimension, the angle between these two vectors is denoted as $\angle(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$. For two sequences $\{a_n\}$, $\{b_n\}$, we let $a_n \ll b_n$ stand for $a_n = O(b_n^\epsilon)$ for some positive $\epsilon < 1$ and let $a_n \succ b_n$ stand for $\lim \frac{b_n}{a_n} = 0$.

3. Consistency of SIR. In order to control the behavior of SIR, we need to impose the following boundedness condition **(A3)** on the predictors' covariance matrix in addition to the tail condition (sub-Gaussian) on their joint distribution. We also need a condition **(A4)** for the central curve.

- **(A3) Boundedness Condition:** $\boldsymbol{x}$ is sub-Gaussian; and there exist positive constants C_1, C_2 such that

$$C_1 \le \lambda_{min}(\mathbf{\Sigma}_{\boldsymbol{x}}) \le \lambda_{max}(\mathbf{\Sigma}_{\boldsymbol{x}}) \le C_2$$

where $\lambda_{min}(\boldsymbol{\Sigma_x})$ and $\lambda_{max}(\boldsymbol{\Sigma_x})$ are the minimal and maximal eigenvalues of $\boldsymbol{\Sigma_x}$ respectively.

- (**A4**) The central curve $\boldsymbol{m}(y) \triangleq \mathbb{E}[\boldsymbol{x}|y]$ has finite fourth moment and is ϑ-sliced stable (defined below) with respect to y and $\boldsymbol{m}(y)$.

DEFINITION 1. *For two positive constants $\gamma_1 < 1 < \gamma_2$, let $\mathcal{A}_H(\gamma_1, \gamma_2)$ be the collection of all the partition $-\infty = a_0 < a_1 < \cdots < a_{H-1} < a_H = \infty$ of $\mathbb{R}$ satisfying that*

$$\frac{\gamma_1}{H} \leq P(a_i \leq y < a_{i+1}) \leq \frac{\gamma_2}{H}.$$

The central curve $\boldsymbol{m}(y) = \mathbb{E}[\boldsymbol{x}|y]$ is called ϑ-sliced stable with respect to y for some $\vartheta > 0$ if there exist positive constants $\gamma_i, i = 1, 2, 3$ such that for any $\boldsymbol{\beta}$ in the central space for any partition in $\mathcal{A}_H(\gamma_1, \gamma_2)$, we have

$$\frac{1}{H}\Big|\sum_{h=0}^{H-1} var(\boldsymbol{\beta}^\tau \boldsymbol{m}(y) \mid a_h \leq y \leq a_{h+1})\Big| \leq \frac{\gamma_3}{H^\vartheta} var(\boldsymbol{\beta}^\tau \boldsymbol{m}(y)). \tag{4}$$

The central curve is sliced stable if it is ϑ-sliced stable for some positive constant ϑ.

REMARK 1. *Note that we only need (4) to hold for all unit vectors in the central space by rescaling. By considering the orthogonal decomposition of $\boldsymbol{\beta}$ in a general space with respect to the central space and its complement, it is easy to see that the sliced stability implies that (4) holds true for all vector $\boldsymbol{\beta}$. In particular, we have the following two useful consequences of the slice-stability.*

i) By choosing $\boldsymbol{\beta}^\tau = (0, \ldots, 0, 1, 0, \ldots, 0)$ with 1 at the k-th position, we have

$$|\sum_{h=0}^{H} var(\boldsymbol{m}(y,k) \mid a_h \leq y \leq a_{h+1})| \leq \gamma_3 H^{1-\vartheta} var(\boldsymbol{m}(y,k)),$$

where $\boldsymbol{m}(y,k)$ is the k-th coordinate of the central curve $\boldsymbol{m}(y)$.

ii) Since equation (4) holds for all unit vector $\boldsymbol{\beta}$, we have

$$\|\sum_{h=0}^{H} var(\boldsymbol{m}(y) \mid a_h \leq y \leq a_{h+1})\|_2 \leq \gamma_3 H^{1-\vartheta} \|var(\boldsymbol{m}(y))\|_2.$$

REMARK 2. *Suppose* $\mathbb{E}[\boldsymbol{m}(y)] = 0$ *and there are* n *samples* $\boldsymbol{m}_i \triangleq \boldsymbol{m}(y_i)$. *Let* $\boldsymbol{m}_{h,i}$ *and* $\overline{\boldsymbol{m}}_{h,\cdot}$ *be defined similarly to* $\boldsymbol{x}_{h,i}$ *and* $\overline{\boldsymbol{x}}_{h,\cdot}$, *respectively. On one hand, we have the classic consistent estimator* $\frac{1}{n}\sum_i \boldsymbol{m}_i\boldsymbol{m}_i^\tau$ *of* $var(\boldsymbol{m}(y))$. *On the other hand, if we expect that the slice-based estimate* $\frac{1}{H}\sum_h \overline{\boldsymbol{m}}_{h,\cdot}\overline{\boldsymbol{m}}_{h,\cdot}^\tau$ *of* $var(\boldsymbol{m}(y))$ *is consistent, we must require that the average loss of variance in each slice to decrease to zero as* H *increases, i.e.,*

$$\frac{1}{H}\sum_h \bar{\boldsymbol{m}}_{h,\cdot}\bar{\boldsymbol{m}}_{h,\cdot}^\tau - \frac{1}{n}\sum_i \boldsymbol{m}_i\boldsymbol{m}_i^\tau = \frac{1}{H}\sum_h \frac{1}{c}\sum_i (\bar{\boldsymbol{m}}_{h,\cdot} - \bar{\boldsymbol{m}}_{h,i})^2 \to 0. \tag{5}$$

In Definition 1, we simply choose the decreasing rate to be a power of H. *It would be easily seen that if* $\boldsymbol{m}$ *is smooth and* y *is compactly supported then* (5) *holds automatically. In this sense, for general curve* $\boldsymbol{m}$ *and random variable* y, *the sliced stability is a condition on smoothness of the central curve* $\boldsymbol{m}$ *and tail distribution of* $\boldsymbol{m}(y)$. *This is not surprised at all, since most work on the consistency of SIR estimate requires some kind of smoothness for the central curve and a tail distribution control for* $\boldsymbol{m}(y)$.

The most popular smoothness and tail condition might be the one proposed by Hsing and Carroll [1992] (later used in Zhu et al. [2006], Zhu and Ng [1995]) in their proof of the consistency of SIR, which is explained below. For $B > 0$ and $n \geq 1$, let $\Pi_n(B)$ be the collection of all the n-point partitions $-B \leq y_{(1)} \leq \cdots \leq y_{(n)} \leq B$ of $[-B, B]$. First, they assumed that the central curve $\boldsymbol{m}(y)$ satisfies the following smoothness condition

$$\lim_{n\to\infty} \sup_{y\in\Pi_n(B)} n^{-1/4} \sum_{i=2}^{n} \|\boldsymbol{m}(y_i) - \boldsymbol{m}(y_{i-1})\|_2 = 0, \forall B > 0.$$

Second, they assumed that for $B_0 > 0$, there exists a non-decreasing function $\widetilde{m}(y)$ on (B_0, ∞), such that

$$\begin{gathered} \widetilde{m}^4(y)P(|Y| > y) \to 0 \text{ as } y \to \infty \\ \|\boldsymbol{m}(y) - \boldsymbol{m}(y')\|_2 \leq |\widetilde{m}(y) - \widetilde{m}(y')| \text{ for } y, y' \in (-\infty, -B_0) \cup (B_0, \infty) \end{gathered} \tag{6}$$

By changing the tail condition (6) to a slightly stronger condition $\mathbb{E}[\widetilde{m}(y)^4] < \infty$, Neykov et al. [2015] proved that the modified condition implies the sliced stability condition. Now, we are ready to state our main results.

THEOREM 1. *Under conditions* (**A1**), (**A2**), (**A3**) *and* (**A4**), *we have*

$$\|\widehat{\boldsymbol{\Lambda}}_H - \boldsymbol{\Lambda}_p\|_2 = O_P\Big(\frac{1}{H^\vartheta} + \frac{H^2 p}{n} + \sqrt{\frac{H^2 p}{n}}\Big). \tag{7}$$

The proof of the theorem is deferred to the Appendix. As a direct consequence of Theorem 1, we observe that if $\rho = \lim_{n\to\infty} \frac{p}{n} = 0$, we may choose $H = \log(n/p)$ such that the right hand side of equation (7) converges to 0. Thus, Theorem 1 implies that $\widehat{\mathbf{\Lambda}}_H$ is a consistent estimate of $\mathbf{\Lambda}_p$ if $\rho = 0$.

REMARK 3 (More on Convergence Rate). *Note that the convergence rate in (7) depends on the choice of H. This may seem not very desirable at the first glance. Since the convergence rate of $\widehat{\mathbf{\Lambda}}_H$ might be different from that of $col(\widehat{\boldsymbol{V}}_H)$, we may expect that the convergence rate of $col(\widehat{\boldsymbol{V}}_H)$ does not depend on the choice of H. In fact, we have*

$$\widehat{\mathbf{\Lambda}}_H - \mathbf{\Lambda}_p = \left(\widehat{\mathbf{\Lambda}}_H - P_{\boldsymbol{V}}\widehat{\mathbf{\Lambda}}_H P_{\boldsymbol{V}}\right) + \left(P_{\boldsymbol{V}}\widehat{\mathbf{\Lambda}}_H P_{\boldsymbol{V}} - \mathbf{\Lambda}_p\right). \tag{8}$$

From the proof of Theorem 1, we can easily check that the first term is of convergence rate $\frac{pH^2}{n} + \sqrt{\frac{pH^2}{n}}$ and the second term is of rate $\frac{1}{H^{\vartheta}}$. Since $P_{\boldsymbol{V}}\widehat{\mathbf{\Lambda}}_H P_{\boldsymbol{V}}$ and $\mathbf{\Lambda}_p$ share the same column space, if we are only interested in estimating $P_{\boldsymbol{V}}$, then the convergence rate of the second term does not matter provided that H is a large enough integer, which may depend on ϑ and $\boldsymbol{\gamma}_3$ but does not depend on n and p. For such an H, if $\mathcal{A}_H(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)$ is non-empty, Theorem 1 and (8) hold for both categorical and continuous response variable Y.

THEOREM 2. *Under conditions* (**A1**), (**A2**), (**A3**), (**A4**) *and assuming that $\rho = \lim \frac{p}{n} = 0$, we have*

$$\|\widehat{\mathbf{\Sigma}}_{\boldsymbol{x}}^{-1}\widehat{\mathbf{\Lambda}}_H - \mathbf{\Sigma}_{\boldsymbol{x}}^{-1}\mathbf{\Lambda}_p\|_2 \to 0 \quad \text{as } n \to \infty$$

with probability converging to one, where $\widehat{\mathbf{\Sigma}}_{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\tau$.

We define the distance $\mathcal{D}(\boldsymbol{V}_1, \boldsymbol{V}_2)$ of two d-dimensional subspaces $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ as the operator norm (or Frobenius norm) of the difference between $P_{\boldsymbol{V}_1}$ and $P_{\boldsymbol{V}_2}$. Simple linear algebra shows that if the $\widetilde{\boldsymbol{\beta}}_i$'s satisfy $\mathbf{\Sigma}_{\boldsymbol{x}}\widetilde{\boldsymbol{\beta}}_i = \lambda_i\mathbf{\Lambda}_p\widetilde{\boldsymbol{\beta}}_i$, then

$$col(\boldsymbol{V}) = span\left\{\widetilde{\boldsymbol{\beta}}_1, \cdots, \widetilde{\boldsymbol{\beta}}_d\right\}.$$

Let $\widehat{\boldsymbol{V}}$ be the matrix formed by the top d generalized eigenvectors of $(\widehat{\mathbf{\Sigma}}_{\boldsymbol{x}}^{-1}, \widehat{\mathbf{\Lambda}}_H)$. Recall that the d-th eigenvalue of $\mathbf{\Lambda}_p$ is assumed to be bounded away from 0. Therefore Theorem 2 implies that $\mathcal{D}(P_{\widehat{\boldsymbol{V}}}, P_{\boldsymbol{V}}) \to 0$ when $\rho = 0$.

We have already shown that the SIR procedure provides us with a consistent estimate of the sufficient dimension reduction space when $\rho = 0$ under

mild conditions. It is then natural to ask: is this condition necessary? Our next theorem gives the answer.

THEOREM 3. *Under conditions* (**A1**), (**A2**), (**A4**) *and assuming that* $\boldsymbol{x} \sim N(0, \mathbf{I}_p)$ *for the single index model*

$$y = f(\boldsymbol{\beta}^\tau \boldsymbol{x}, \epsilon),$$

we have:

(i) When $\rho = \lim \frac{p}{n} \in (0, \infty)$, $\|\widehat{\boldsymbol{\Lambda}}_H - \boldsymbol{\Lambda}_p\|_2$, *as a function of* ρ, *is dominated by* $\sqrt{\rho} \vee \rho$ *when* $H, n \to \infty$;

(ii) Let $\widehat{\boldsymbol{\beta}}$ *be the principal eigenvector of the SIR estimator* $\widehat{\boldsymbol{\Lambda}}_H$. *If* $\rho = \lim \frac{p}{n} > 0$, *then there exists a positive constant* $c(\rho) > 0$ *such that*

$$\liminf_{n \to \infty} \mathbb{E} \angle(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}}) > c(\rho)$$

with probability converges to one.

We illustrate this result via a numerical study of the linear model

$$y = \boldsymbol{x}^\tau \boldsymbol{\beta} + \epsilon \text{ where } \boldsymbol{\beta}^\tau = (1, 0, \cdots, 0), \boldsymbol{x} \sim N(0, \mathbf{I}_p), \epsilon \sim N(0, 1). \tag{9}$$

Figure 1 shows how $\mathbb{E} \angle(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}})$ is related to the dimension p for fixed ratio $\rho = \frac{p}{n}$ (taking values in $\{.1, .3, .7, 1, 2, 4\}$), where $\boldsymbol{\beta}$ is estimated by the SIR with the slice number $H = 10$. For each p, $\mathbb{E} \angle(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}})$ is calculated based on 100 iterations. It is seen that this expected angle converges to a positive number when the ratio ρ is non-zero. In Figure 2, we have plotted the $\mathbb{E} \angle(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}})$ against the ratio $\rho = \frac{p}{n}$, varying between 0.01 and 4 with an increment of 0.01. The sample size n is 200 and the slice number H is 10. It is seen that the expected angle decreases to zero as ρ approaches zero, and increases monotonically when ρ increases.

Results in this section have shown that there is a phase transition phenomenon of the SIR procedure. That is, the estimate of the dimension reduction space is consistent if and only if the ratio $\rho = \lim \frac{p}{n} = 0$. This provides a theoretical justification of imposing additional structure assumption such as sparsity in high dimension.

4. SIR in ultra-high dimension. As we have shown in Section 3, the SIR estimator fails to be consistent if $\rho = \lim \frac{p}{n} \neq 0$. Hence, when $p \gg n$, some structural assumptions are necessary for getting a consistent estimate of the central space. In this paper, we assume that both the loadings of all the

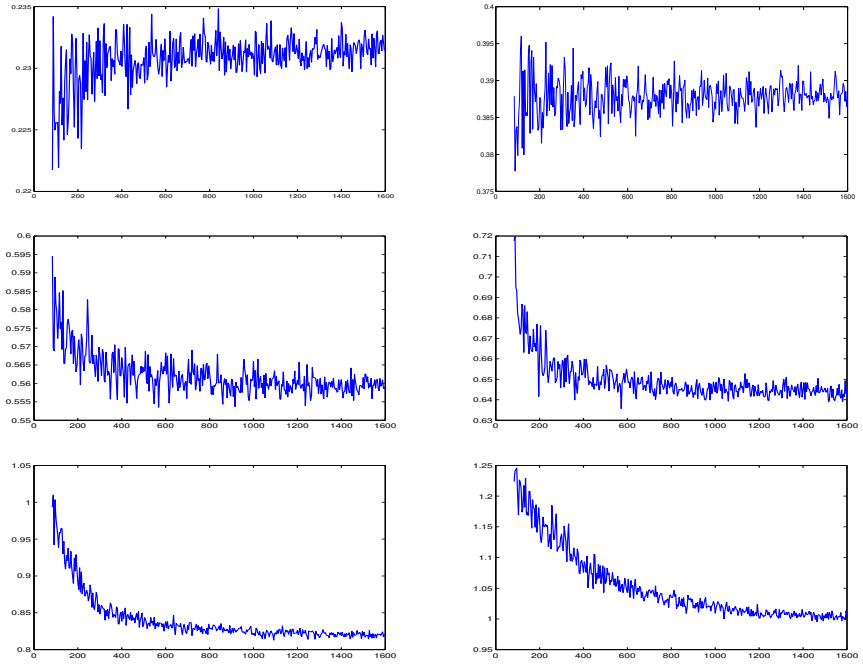

Fig 1: Numerical approximations of $\mathbb{E}\angle(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ for model (9) as a function of dimension p for $\rho = .1,\ .3,\ .7,\ 1,\ 2,$ and 4, respectively (up left, up right, middle left, middle right, lower left, lower right), where $\widehat{\boldsymbol{\beta}}$ is estimated by SIR.

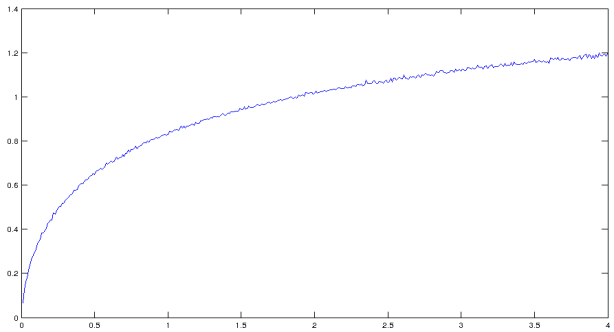

Fig 2: The relationship of $\mathbb{E}\angle(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}})$ and the ratio p/n where $\widehat{\boldsymbol{\beta}}$ is estimated by SIR.

directions $\boldsymbol{\beta}_j$'s and the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ are sparse. Other structural assumptions will be studied in our future work. For $\boldsymbol{\beta}_i$'s, we impose the following prevalent sparsity condition.

- **(A5)** $s = |\mathcal{S}| \ll p$ where $\mathcal{S} = \left\{ i \mid \boldsymbol{\beta}_j(i) \neq 0 \text{ for some } j,\ 1 \leq j \leq d \right\}$ and $|\mathcal{S}|$ is the number of elements in the set $\mathcal{S}$.

For $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, the following class of covariance matrices has been introduced in Bickel and Levina [2008] (see also Cai et al. [2010]).

$$\mathcal{U}(\epsilon_0, \alpha, C) = \Big\{ \boldsymbol{\Sigma}_{\boldsymbol{x}} : \max_j \sum_i \{|\sigma_{i,j}| : |i-j| > l\} \leq C l^{-\alpha} \text{ for all } l > 0, \\ \text{and } 0 < \epsilon_0 \leq \lambda_{min}(\boldsymbol{\Sigma}_{\boldsymbol{x}}) \leq \lambda_{max}(\boldsymbol{\Sigma}_{\boldsymbol{x}}) \leq \frac{1}{\epsilon_0} \Big\}.$$

In this paper, to simplify the notations and arguments, we choose a slightly stronger condition.

- **(A6)** $\boldsymbol{\Sigma}_{\boldsymbol{x}} \in \mathcal{U}(\epsilon_0, \alpha, C)$ and $\max_{1 \leq i \leq p} r_i$ is bounded where r_i is the number of non-zero elements in the i-th row of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$.

Let $\mathcal{T} = \left\{ k \mid var\left(\mathbb{E}[\boldsymbol{x}(k)|y]\right) \neq 0 \right\}$. If $k \in \mathcal{T}$, there exists $\boldsymbol{\eta} \in col(\boldsymbol{\Lambda})$ such that $\boldsymbol{\eta}(k) \neq 0$. Since we have (3):

$$\boldsymbol{\Sigma}_{\boldsymbol{x}} col(\boldsymbol{V}) = col(\boldsymbol{\Lambda}),$$

there exits a $\boldsymbol{\beta} \in col(\boldsymbol{V})$ such that $\boldsymbol{\eta} = \Sigma_{\boldsymbol{x}}\boldsymbol{\beta}$. Thus if $k \in \mathcal{T}$, then $k \in supp(\Sigma_{\boldsymbol{x}}\boldsymbol{\beta})$ for some $\boldsymbol{\beta} \in col(\boldsymbol{V})$. In particular, with the above sparsity assumptions **(A5)** and **(A6)** , we have $|\mathcal{T}| \leq s \max_{1 \leq i \leq p} r_i = O(s)$.[2] Note that our goal here is to recover the column space $col(V)$ rather than $\mathcal{S}$. Indeed, we are not able to consistently recover $\mathcal{S}$ unless for the trivial case. The key for recovering $cov(V)$ is to consistently recovering the set $\mathcal{T}$.

At the population level, $var(\mathbb{E}(\boldsymbol{x}(k))|y)$ can separate $\mathcal{T}$ from $\mathcal{T}^c$. When there are only finite samples, we use

$$var_H(\boldsymbol{x}(k)) = \frac{1}{H} \sum_{h=1}^{H} \bar{\boldsymbol{x}}_{h,\cdot}(k)^2 \tag{10}$$

as an estimate of $var(\mathbb{E}(\boldsymbol{x}(k))|y)$. These are the diagonal elements of the matrix $\widehat{\boldsymbol{\Lambda}}_H$. Note that these quantities depend on the sliced sample means, which are neither independent nor identically distributed. Thus, the usual

[2]We could introduce $\xi = \max_{1 \leq i \leq p} r_i$, then $|\mathcal{T}| \leq s\xi$. The arguments below still work, except we might need $s\xi = o(p)$.

concentration inequalities for χ^2 are no longer applicable. We need extra efforts to get the concentration inequalities; this concentration result is one of the main technical contributions of this article, and can be further generalized.

REMARK 4. *The link function $f(\ )$ is not involved explicitly in the definition of $var_H(\boldsymbol{x}(k))$, and only the order statistics of the response is required. This nonparametric characteristic of the method is of particular interest to us and will be further investigated in a future research. Screening statistics inspired by the sliced inverse regression idea have been proposed in various formats, such as those in Jiang and Liu [2014], Zhu et al. [2012] and Cui et al. [2015].*

With the quantities $var_H(\mathbb{E}[\boldsymbol{x}(k)|y])$, we define the inclusion set $\mathcal{I}_p(t)$ and the exclusion set $\mathcal{E}_p(t)$ below, which depend on a thresholding value t:

$$\mathcal{I}_p(t) = \left\{ \ k \ | var_H(\boldsymbol{x}(k)) > t \ \right\} \text{ and } \mathcal{E}_p(t) = \left\{ \ k \ | var_H(\boldsymbol{x}(k)) \leq t \ \right\}.$$

Note that $\mathcal{I}_p(t)$ can be viewed as an estimate of $\mathcal{T}$ and is thus also denoted by $\widehat{\mathcal{T}}$. After reducing the dimension to a level such as p/n is sufficiently small, the SIR estimator $\widehat{\boldsymbol{\Lambda}}^{\widehat{\mathcal{T}},\widehat{\mathcal{T}}}$ is a consistent estimate of $\boldsymbol{\Lambda}^{\mathcal{T},\mathcal{T}}$. Let $\widehat{\boldsymbol{V}}^{\widehat{\mathcal{T}}}$ be the matrix formed by the top d eigenvectors of $\widehat{\boldsymbol{\Lambda}}^{\widehat{\mathcal{T}},\widehat{\mathcal{T}}}$. We then use $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} col(e(\widehat{\boldsymbol{V}}^{\widehat{\mathcal{T}}}))$ to estimate the central space $col(\boldsymbol{V})$, where $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}$ is a consistent estimate of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$. Estimating the covariance matrix and precision matrix in high dimension is a challenging problem by itself and is not a main focus of this article. We employ the methods of Bickel and Levina [2008] to solve it. In summary, we propose the following **D**iagonal **T**hresholding screening **SIR** (DT-SIR) algorithm:

Algorithm 1 DT-SIR

1. Calculate $var_H(\boldsymbol{x}(k))$ according (10) for $k = 1, 2, \cdots, p$;
2. Let $\widehat{\mathcal{T}} = \left\{ \ k \mid var_H(\boldsymbol{x}(k)) > t \ \right\}$ for an appropriate t ;
3. Let $\widehat{\boldsymbol{\Lambda}}_H^{\widehat{\mathcal{T}},\widehat{\mathcal{T}}}$ be the SIR estimator of the conditional covariance matrix for the data $(y, \boldsymbol{x}^{-,\widehat{\mathcal{T}}})$ according to equation (2);
4. Let $\widehat{\boldsymbol{V}}^{\widehat{\mathcal{T}}}$ be the matrix formed by the top d eigenvectors of $\widehat{\boldsymbol{\Lambda}}^{\widehat{\mathcal{T}},\widehat{\mathcal{T}}}$;
5. $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} col\left(e\left(\hat{v}^{\widehat{\mathcal{T}}}\right)\right)$ is the estimate of $col(\boldsymbol{V})$

A practical way to choose an appropriate t in step 2 will be presented in Section 5 . To ensure theoretical properties, we need an assumption on the signal strength:

- **(S1)** $\exists\ C > 0$ and $\omega > 0$ such that $var(\mathbb{E}[\boldsymbol{x}(k)|y]) > Cs^{-\omega}$ when $\mathbb{E}[\boldsymbol{x}(k)|y]$ is not a constant.

THEOREM 4. *Under conditions* **(A1)** $-$ **(A6)** *and* **(S1)**, *and let* $t = as^{-\omega}$ *for some constant* $a > 0$ *such that* $t < \frac{1}{2}var(m(y,k),\ \forall k \in \mathcal{T}$, *we have*

i) $\mathcal{T}^c \subset \mathcal{E}_p$ *holds with probability at least*

$$1 - C_1 \exp\left(-C_2 \frac{n}{H^2 s^\omega} + C_3 \log(H) + \log(p-s)\right); \tag{11}$$

ii) $\mathcal{T} \subset \mathcal{I}_p$ *holds with probability at least*

$$1 - C_4 \exp\left(-C_5 \frac{n}{H^2 s^\omega} + C_6 \log(H) + \log(s)\right), \tag{12}$$

for some positive constants $C_1, \cdots, C_6$.

This theorem has a simple implication. If $\frac{n}{s^\omega} \succ \log(p) + \log(s)$, we may choose $H = \log(\frac{n}{s^\omega \log(p)})$, so that

$$\frac{n}{H^2 s^\omega} \succ \log(p) + \log(H) + \log(s).$$

Thus , we know $\mathcal{T} = \mathcal{I}_p$ with probability converging to one. Next, we have results for the consistency of DT-SIR.

THEOREM 5. *Under the same assumptions and choosing the same* t *as Theorem 4, if* $\frac{n}{s^\omega} \succ \log(p) + \log(s)$, *we have*

$$\|e(\widehat{\boldsymbol{\Lambda}}_H^{\widehat{\mathcal{T}},\widehat{\mathcal{T}}}) - \boldsymbol{\Lambda}_p\|_2 \to 0 \ \text{as}\ n \to \infty$$

with probability converging to one, where $\widehat{\mathcal{T}} = \mathcal{I}(t)$ *and* $H = \log(\frac{n}{s^\omega \log(p)})$.

THEOREM 6. *Let* $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}$ *be the estimator of co-variance matrix from Bickel and Levina [2008]. Under the same assumptions of Theorem 5, we have*

$$\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} e(\widehat{\boldsymbol{\Lambda}}_H^{\widehat{\mathcal{T}},\widehat{\mathcal{T}}}) - \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Lambda}_p\|_2 \to 0 \ \text{as}\ n \to \infty$$

with probability converging to one.

5. Simulation Studies. We consider the following settings in generating the design matrix $\boldsymbol{x}$ and the response y. In Settings I-III, each row of $\boldsymbol{x}$ is independently sampled from $N(\mathbf{0}, \mathbf{I})$.

- **Setting I.** $y_i = \sin(x_{i1} + x_{i2}) + \exp(x_{i3} + x_{i4}) + 0.5 * \epsilon_i$, where $\epsilon_i \overset{\text{iid}}{\sim} N(0, 1)$;
- **Setting II.** $y_i = \sum_{j=1}^{7} x_{ij} * \exp(x_{i8} + x_{i9}) + 0.5 * \epsilon_i$ where $\epsilon_i \overset{\text{iid}}{\sim} N(0, 1)$;
- **Setting III.** $y_i = \sum_{j=1}^{10} x_{ij} * \exp(\sum_{i=11}^{20} x_{ij}) + \epsilon_i$ where $\epsilon_i \overset{\text{iid}}{\sim} N(0, 1)$;

In Settings IV to VI, each row of $\boldsymbol{x}$ is independently sampled from $N(\mathbf{0}, \boldsymbol{\Sigma})$.

- **Setting IV.** $y_i = (x_{i1} + x_{i2} + x_{i3})^3/2 + 0.5 * \epsilon_i$, where $\epsilon_i \overset{\text{iid}}{\sim} N(0, 1)$ and $\boldsymbol{\Sigma} = (\sigma_{ij})$ is tri-diagonal with $\sigma_{ii} = 1$, $\sigma_{i,i+1} = \sigma_{i+1,i} = \rho$ and $\sigma_{i,i+2} = \sigma_{i+2,i} = \rho^2$;
- **Setting V.** $y_i = \sum_{j=1}^{7} x_{ij} * \exp(x_{i8} + x_{i9}) + \epsilon_i$, where $\epsilon_i \overset{\text{iid}}{\sim} N(0, 1)$, and $\boldsymbol{\Sigma} = \boldsymbol{B} \otimes \mathbf{I}_{p/10}$ with $\boldsymbol{B} = (b_{ij})_{1 \le i \le 10, 1 \le j \le 10}$ given as $b_{ij} = \rho^{|i-j|}$;
- **Setting VI.** Assume the same setting as in Setting V except that $\boldsymbol{\Sigma} = (\sigma_{ij})$ is tri-diagonal with $\sigma_{ii} = 1, \sigma_{i,i+1} = \sigma_{i+1,i} = \rho$ and $\sigma_{i,i+2} = \sigma_{i+2,i} = \rho^2$.
- **Setting VII.** Assume the same setting as in Setting V except that $\boldsymbol{\Sigma} = (\sigma_{ij})$ is given as $\sigma_{ij} = \rho^{|i-j|}$.

DT-SIR first screens all the predictors according to the statistic $var_{H,c}(\boldsymbol{x}(k))$, which requires a tuning parameter t. We chose t by using an auxiliary variable method based on an idea first proposed by Luo et al. [2006] and extended by Wu et al. [2007] and Zhu et al. [2011]. In our setting, for a given sample $(y_i, \boldsymbol{x}_i)$, we generate $\boldsymbol{z}_i \sim N(\mathbf{0}, \mathbf{I}_{p'})$ where p' is sufficiently large and chosen as p in our simulation studies. It is known that $\boldsymbol{y}$ and $\boldsymbol{z}$ are independent. The threshold t can be chosen as

$$\hat{t} = \max_{1 \le k \le p'} \{var_{H,c}(\boldsymbol{z}(k)).\}$$

In DT-SIR, when $n > 1000$, H is chosen as 20; when $n \le 1000$, H is chosen as 10 in the screening step and 20 in the SIR step.

We also consider the following alternative methods in the screening step: Sure Independent Ranking and Screening (SIRS) in Zhu et al. [2011], SIR for variable selection via Inverse modeling (SIRI) in Jiang and Liu [2014], and trace pursuit in Yu et al. [2016]. As a comparison, we also considered two screening methods that are not based on the sliced regression: Distance correlation in Székely et al. [2007] and SURE independence Fan and Lv [2008]. For SIRS, the threshold is chosen according to the auxiliary statistic

(2.9) of Zhu et al. [2011]. For SIRI, the predictors are chosen according to 10-fold cross validation. The threshold values $\bar{c}^{SIR}$ and $\underline{c}^{SIR}$ are chosen as the 10-th and 5-th quantile of a weighted χ^2 distribution given in Theorem 3.1 of Yu et al. [2016]. In both SURE and DC screening, the top $\lfloor \gamma n \rfloor$ where $\gamma = 0.01$ are kept for subsequent analyses.

After the screening step, similar to DT-SIR, we then applied the SIR algorithm (steps 3-5 of DT-SIR) to estimate $col(\boldsymbol{V})$. These alternative methods are denoted as SIRS-SIR, SIRI-SIR, SURE-SIR, DC-SIR, and TP-SIR, respectively, in the following discussions. Another method that we compared with is the sparse SIR, abbreviated as SpSIR, proposed in Li [2007]. After obtaining an estimator $col(\widehat{\boldsymbol{V}})$, we calculate $\mathcal{D}(P_{col(\hat{V})}, P_{col(\boldsymbol{V})})$ as a measure of the estimation error. We replicate this step 100 times, and calculate the average distance for the estimation result from each method and report these numbers in Table 1-3. For each setting, the average distance of the optimal method is highlighted using bold fonts. We further run a two-sample T-test to test if the actual estimation error of each method is significantly different from that of the best method for that example at 1% level of significance.

Under all settings, the average distance obtained by DT-SIR was much smaller than that obtained by SpSIR and SURE-SIR. The p-values for comparing DT-SIR and SpSIR/SURE-SIR are all significant at the 0.01 level. When $p \geq n$, the sparse SIR completely failed because the average distance of the estimated space to the true space is $\sqrt{2d}$, indicating that the space estimated by sparse SIR is orthogonal to the true space spanned by β.

Under settings II-IV, DT-SIR performed either the best or not significantly worse than the best method. For all other cases, DT-SIR performed the best except for a few cases: Setting I when $n = 500, p = 1000$, setting V when $n = 500, p = 6000$, setting VI when $n = 500, p = 6000$, and setting VII when $n = 1000, p = 1000$.

When $p = 6000, n = 500$, both DT-SIR and SIRI-SIR were the winners. Under Setting III, DT-SIR performed better than SIRI-SIR; under settings V and VI, SIRI-SIR performed better than DT-SIR; under other settings, these two methods were comparable.

To graphically show the performance of various methods, we consider setting IV with $d = 1$. Consider two cases when $(n, p) = (2000, 1000)$ and $(n, p) = (500, 100)$. We calculated the estimated directions $\hat{\boldsymbol{\beta}}$ using various methods and computed the angle between $< \hat{\boldsymbol{\beta}} >$ and $< \boldsymbol{\beta} >$. We replicate this step 100 times to calculate the average angles for each method. The results are displayed in Figure 3, which shows clearly that DT-SIR performed better than its competitors.

Additionally, DT-SIR is computationally efficient. To show this, we re-

TABLE 1

The average distance of the space estimated by each of the 7 methods tested to the true space $col(\boldsymbol{V})$ under various settings with $p = 1000$. The boldfaced number in each row represents the best result for that simulation scenario, and the "" in cells represents that the p-value of the two-sample T-test comparing the estimation error of the corresponding method with that of the best method is less than 0.01.*

	n	DT-SIR	SIRI-SIR	SIRS-SIR	SpSIR	SURE-SIR	DC-SIR	TP-SIR
I	500	0.655(*)	0.751(*)	**0.492**	2(*)	1.39(*)	0.731(*)	1.18(*)
	1000	**0.3**	0.431(*)	0.309	2(*)	1.29(*)	0.632(*)	0.94(*)
	2000	**0.221**	0.341(*)	0.226	1.58(*)	1.04(*)	0.655(*)	0.784(*)
	3000	0.167	0.245(*)	**0.149**	1.48(*)	0.816(*)	0.641(*)	0.713(*)
II	500	0.383	0.396	**0.371**	2(*)	1.64(*)	1.08(*)	0.389
	1000	0.235	**0.227**	0.256	2(*)	1.36(*)	0.266(*)	0.318(*)
	2000	0.161	**0.157**	0.189(*)	1.25(*)	1.25(*)	0.387(*)	0.264(*)
	3000	0.134	**0.129**	0.153(*)	0.975(*)	1.12(*)	0.404(*)	0.23(*)
III	500	1.15	1.48(*)	1.38(*)	2(*)	1.97(*)	1.85(*)	**1.13**
	1000	**0.426**	0.974(*)	0.596(*)	2(*)	1.94(*)	1.57(*)	0.429
	2000	**0.263**	0.403(*)	0.29(*)	1.33(*)	1.89(*)	0.996(*)	0.338(*)
	3000	**0.214**	0.297	0.238(*)	1.06(*)	1.82(*)	0.475(*)	0.299(*)
IV	500	0.263	**0.257**	0.333	1.41(*)	0.335(*)	0.334(*)	0.332(*)
	1000	**0.219**	0.447(*)	0.25	1.41(*)	0.436(*)	0.459(*)	0.469(*)
	2000	**0.161**	0.4(*)	0.196(*)	0.42(*)	0.442(*)	0.469(*)	0.452(*)
	3000	**0.134**	0.377(*)	0.177(*)	0.297(*)	0.43(*)	0.458(*)	0.438(*)
V	500	0.546	**0.529**	0.562	2(*)	1.62(*)	1.24(*)	1.09(*)
	1000	0.401	0.463(*)	0.514(*)	2(*)	1.15(*)	**0.367**	0.615(*)
	2000	**0.288**	0.418(*)	0.341(*)	1.51(*)	0.926(*)	0.569(*)	0.54(*)
	3000	**0.249**	0.399(*)	0.284(*)	1.24(*)	0.691(*)	0.597(*)	0.511(*)
VI	500	0.568	**0.535**	0.566	2(*)	1.64(*)	1.24(*)	1.08(*)
	1000	0.427	0.524(*)	0.548(*)	2(*)	1.22(*)	**0.39**	0.641(*)
	2000	**0.311**	0.469(*)	0.351(*)	1.51(*)	0.927(*)	0.598(*)	0.583(*)
	3000	**0.265**	0.456(*)	0.307(*)	1.25(*)	0.807(*)	0.622(*)	0.56(*)
VII	500	0.556	**0.534**	0.585(*)	2(*)	1.66(*)	1.26(*)	1.11(*)
	1000	0.436(*)	0.528(*)	0.545(*)	2(*)	1.22(*)	**0.39**	0.643(*)
	2000	**0.303**	0.465(*)	0.358(*)	1.51(*)	0.747(*)	0.589(*)	0.579(*)
	3000	**0.258**	0.468(*)	0.319(*)	1.25(*)	0.698(*)	0.63(*)	0.558(*)

port the computing time for one replication under Setting II for various pairs of (n, p) in Table 4. All computations were done on a computer with Intel Xeon(R) E5-1620 CPU@3.70GHz and 16GB memory. It is clearly seen that DT-SIR performed as fast as SURE-SIR, and both were much faster than other competitors. Consider the case when $p = 3000, n = 2000$. The computation time of DT-SIR was only 30 seconds; while that for DC-SIR was 21 minutes and 38 seconds, and the that for TP-SIR was 6 minutes and 17 seconds.

TABLE 2
The average distance of the space estimated by each of the 7 methods we tested to the true space $col(\boldsymbol{V})$ under various settings with $n = 2000$.

	p	DT-SIR	SIRI-SIR	SIRS-SIR	SpSIR	SURE-SIR	DC-SIR	TP-SIR
I	500	0.213	0.312(*)	**0.206**	1.44(*)	0.903(*)	0.629(*)	0.772(*)
	1000	**0.221**	0.341(*)	0.226	1.58(*)	1.04(*)	0.655(*)	0.784(*)
	2000	0.241	0.29	**0.214**	2(*)	1.07(*)	0.677(*)	0.793(*)
	3000	0.23	0.278	**0.218**	2(*)	1.17(*)	0.683(*)	0.797(*)
II	500	0.163	**0.16**	0.19(*)	0.83(*)	1.22(*)	0.369(*)	0.26(*)
	1000	0.161	**0.157**	0.189(*)	1.25(*)	1.25(*)	0.387(*)	0.264(*)
	2000	0.172	**0.159**	0.196(*)	2(*)	1.23(*)	0.404(*)	0.259(*)
	3000	0.164	**0.158**	0.199(*)	2(*)	1.3(*)	0.414(*)	0.261(*)
III	500	**0.272**	0.353	0.29(*)	0.916(*)	1.84(*)	0.846(*)	0.341(*)
	1000	**0.263**	0.403(*)	0.29(*)	1.33(*)	1.89(*)	0.996(*)	0.338(*)
	2000	**0.262**	0.368	0.285(*)	2(*)	1.92(*)	0.98(*)	0.339(*)
	3000	**0.269**	0.344	0.291(*)	2(*)	1.93(*)	1.09(*)	0.339(*)
IV	500	**0.145**	0.409(*)	0.182(*)	0.248(*)	0.406(*)	0.433(*)	0.438(*)
	1000	**0.161**	0.4(*)	0.196(*)	0.42(*)	0.442(*)	0.469(*)	0.452(*)
	2000	**0.16**	0.395(*)	0.198(*)	1.41(*)	0.472(*)	0.506(*)	0.447(*)
	3000	**0.15**	0.395(*)	0.216(*)	1.41(*)	0.49(*)	0.527(*)	0.447(*)
V	500	**0.272**	0.434(*)	0.353(*)	1.09(*)	0.876(*)	0.547(*)	0.539(*)
	1000	**0.288**	0.418(*)	0.341(*)	1.51(*)	0.926(*)	0.569(*)	0.54(*)
	2000	**0.289**	0.418(*)	0.351(*)	2(*)	0.868(*)	0.596(*)	0.537(*)
	3000	**0.3**	0.417(*)	0.372(*)	2(*)	0.968(*)	0.605(*)	0.544(*)
VI	500	**0.307**	0.479(*)	0.368(*)	1.1(*)	0.858(*)	0.566(*)	0.583(*)
	1000	**0.311**	0.469(*)	0.351(*)	1.51(*)	0.927(*)	0.598(*)	0.583(*)
	2000	**0.309**	0.461(*)	0.399(*)	2(*)	1.08(*)	0.617(*)	0.585(*)
	3000	**0.31**	0.46(*)	0.408(*)	2(*)	1(*)	0.638(*)	0.587(*)
VII	500	**0.299**	0.482(*)	0.343(*)	1.09(*)	0.818(*)	0.564(*)	0.583(*)
	1000	**0.303**	0.465(*)	0.358(*)	1.51(*)	0.747(*)	0.589(*)	0.579(*)
	2000	**0.309**	0.455(*)	0.383(*)	2(*)	0.966(*)	0.622(*)	0.578(*)
	3000	**0.308**	0.46(*)	0.357(*)	2(*)	0.858(*)	0.626(*)	0.58(*)

TABLE 3
The average distance of the space estimated by each of the 7 methods tested to the true space $col(\boldsymbol{V})$ under various settings with $n = 500$ and $p = 6000$.

	DT-SIR	SIRI-SIR	SIRS-SIR	SpSIR	SURE-SIR	DC-SIR	TP-SIR
I	0.694	0.631	**0.606**	2(*)	1.43(*)	0.97(*)	1.19(*)
II	0.446	0.462	0.414	2(*)	1.74(*)	1.08(*)	**0.4**
III	**1.35**	1.56(*)	1.56(*)	2(*)	1.99(*)	1.88(*)	1.37
IV	0.163	**0.122**	0.245(*)	1.41(*)	0.27(*)	0.305(*)	0.195(*)
V	0.481(*)	**0.431**	0.486(*)	2(*)	1.62(*)	1.1(*)	0.995(*)
VI	0.463(*)	**0.423**	0.494(*)	2(*)	1.62(*)	1.11(*)	0.999(*)
VII	0.44	**0.412**	0.477(*)	2(*)	1.61(*)	1.1(*)	1.03(*)

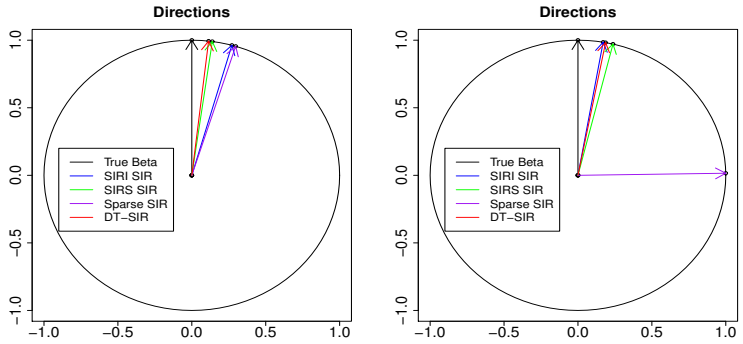

Fig 3: Simulated value of $E\angle(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ for the various methods. Left panel: $(n, p) = (2000, 1000)$; Right panel: $(n, p) = (500, 1000)$.

TABLE 4
Comparison of computing time under setting II.

	DT-SIR	SIRI-SIR	SIRS-SIR	SpSIR	SURE-SIR	DC-SIR	TP-SIR
n	p=1000						
500	1”	1’12”	7”	11”	1”	24”	29”
1000	2”	2’2”	20”	11”	1”	1’52”	1’2”
2000	3”	3’27”	1’14”	13”	2”	7’38”	2’18”
3000	4”	4’59”	2’45”	15”	3”	6’51”	3’7”
p	n=2000						
500	1”	2’48”	35”	2”	1”	3’46”	1’7”
1000	3”	3’27”	1’14”	13”	2”	7’38”	2’18”
2000	12”	4’55”	2’35”	1’39”	12”	14’24”	3’22”
3000	30”	6’0”	4’10”	5’19”	30”	21’38”	6’17”

6. Conclusion. When the dimension p diverges to infinity, classical statistical procedures often fail unless additional structures such as sparsity conditions are imposed. Understanding boundary conditions of a statistical procedure provides us theoretical justification and practical guidance for our modeling efforts. In this article, we provide a new framework to show that $\rho = \lim \frac{p}{n}$ is the phase transition parameter for the SIR procedure. Under certain conditions, it is shown that the SIR estimator is consistent if and only if $\rho = 0$. When $\rho > 0$, where the original SIR fails to be consistent, we propose a two-stage method, DT-SIR for variable screening and selection in ultra-high dimension situations and show that the method is consistent. We have used simulated examples to demonstrate the advantages of DT-SIR

compared to its competitors. This method is computationally fast and can be easily implemented for large data sets.

Appendices

In the following two sections we offer some details about our theoretical derivations, but some more tedious intermediate steps (organized as Lemmas 6-21) are deferred to the Supplemental Document to this article, which is available on line.

A. The Key Lemma. The following lemma plays an important role in developing the high dimensional theory for sliced inverse regression. The proof of this key lemma is lengthy and technical. It will be helpful to keep in mind that H and ν (if they are not constants) grow at very slow rate compared with c and n (e.g., polynomial of $\log(n)$). Let $\boldsymbol{m}(y) = \mathbb{E}[\boldsymbol{x}|y]$, and $\boldsymbol{x} = \boldsymbol{m}(y) + \epsilon$. Notations $\boldsymbol{m}_{h,j}$, $\overline{\boldsymbol{m}}_{h,\cdot}$, $\overline{\overline{\boldsymbol{m}}}$, and $\boldsymbol{\epsilon}_{h,j}$, $\overline{\boldsymbol{\epsilon}}_{h,\cdot}$, $\overline{\overline{\boldsymbol{\epsilon}}}$ are similarly defined as $\boldsymbol{x}_{h,j}$, $\overline{\boldsymbol{x}}_{h,\cdot}$ and $\overline{\overline{\boldsymbol{x}}}$ that were introduced before.

LEMMA 1. *Let $\boldsymbol{x} \in \mathbb{R}^p$ be a sub-Gaussian random variable which is upper exponentially bounded by K (see Definition 4). For any unit vector $\boldsymbol{\beta} \in \mathbb{R}^p$, let $\boldsymbol{x}(\boldsymbol{\beta}) = \langle \boldsymbol{x}, \boldsymbol{\beta} \rangle$ and $\boldsymbol{m}(\boldsymbol{\beta}) = \langle \boldsymbol{m}, \boldsymbol{\beta} \rangle = \mathbb{E}[\boldsymbol{x}(\boldsymbol{\beta}) \mid y]$, we have the following:*

i) If $var(\boldsymbol{m}(\boldsymbol{\beta})) = 0$, there exists positive constants C_1, C_2 and C_3 such that for any $b = O(1)$ and sufficiently large H, we have

$$\mathbb{P}(var_H(\boldsymbol{x}(\boldsymbol{\beta})) > b) \leq C_1 \exp\left(-C_2 \frac{nb}{H^2} + C_3 \log(H)\right).$$

ii) If $var(\boldsymbol{m}(\boldsymbol{\beta})) \neq 0$, there exists positive constants C_1, C_2 and C_3 such that, for any $\nu > 1$, we have

$$|var_H(\boldsymbol{x}(\boldsymbol{\beta})) - var(\boldsymbol{m}(\boldsymbol{\beta}))| \geq \frac{1}{2\nu} var(\boldsymbol{m}(\boldsymbol{\beta}))$$

with probability at most

$$C_1 \exp\left(-C_2 \frac{n\ var(\boldsymbol{m}(\boldsymbol{\beta}))}{H^2 \nu^2} + C_3 \log(H)\right).$$

where we choose H such that $H^\vartheta > C_4 \nu$ for some sufficiently large constant C_4.

A.1. *Proof of Lemma 1 i)* If $\boldsymbol{m}(\boldsymbol{\beta}) = 0$ (or equivalently $var(\boldsymbol{m}(\boldsymbol{\beta})) = 0$), since

$$\begin{aligned}\bar{\boldsymbol{\epsilon}}_{h,\cdot}(\boldsymbol{\beta})^2 =& \left(\frac{c-1}{c}\frac{1}{c-1}\sum_{i=1}^{c-1}\boldsymbol{\epsilon}_{h,i}(\boldsymbol{\beta}) + \frac{1}{c}\boldsymbol{\epsilon}_{h,c}(\boldsymbol{\beta})\right)^2 \\ \leq& 2\left(\frac{1}{c-1}\sum_{i=1}^{c-1}\boldsymbol{\epsilon}_{h,i}(\boldsymbol{\beta})\right)^2 + 2\left(\frac{1}{c}\boldsymbol{\epsilon}_{h,c}(\boldsymbol{\beta})\right)^2\end{aligned}$$

for $h = 1, ..., H-1$ and $\bar{\boldsymbol{\epsilon}}_{H,\cdot}(\boldsymbol{\beta}) = \frac{1}{c}\sum_{i=1}^{c}\boldsymbol{\epsilon}_{H,i}(\boldsymbol{\beta})$, we have

$$\begin{aligned}& var_H(\boldsymbol{x}(\boldsymbol{\beta})) - var(\boldsymbol{m}(\boldsymbol{\beta})) \\ =& \frac{1}{H}\sum_{h}^{H-1}\bar{\boldsymbol{\epsilon}}_{h,\cdot}(\boldsymbol{\beta})^2 + \frac{1}{H}\bar{\boldsymbol{\epsilon}}_{H,\cdot}(\boldsymbol{\beta})^2 \\ \leq& \frac{2}{H}\left(\sum_{h}^{H-1}\left(\frac{1}{c-1}\sum_{i=1}^{c-1}\boldsymbol{\epsilon}_{h,i}(\boldsymbol{\beta})\right)^2 + \bar{\boldsymbol{\epsilon}}_{H,\cdot}(\boldsymbol{\beta})^2\right) + \frac{2}{Hc^2}\sum_{h}^{H-1}\boldsymbol{\epsilon}_{h,c}(\boldsymbol{\beta})^2 \\ \triangleq& 2I + 2II.\end{aligned}$$

Thus

$$\mathbb{P}(var_H(\boldsymbol{x}(\boldsymbol{\beta})) > b) \leq \mathbb{P}(I > b/4) + \mathbb{P}(II > b/4). \tag{13}$$

Lemma 17 (iii) in Supplement implies that

$$\mathbb{P}\left(\boldsymbol{\epsilon}(\boldsymbol{\beta})|_{y\in S_h} > t\right) \leq CH\exp\left(-\frac{t^2}{K^2}\right)$$

for some positive constant C. Since $\mathbb{E}[\boldsymbol{x}(\boldsymbol{\beta})|y] = 0$,we have $\mathbb{E}[\boldsymbol{x}(\boldsymbol{\beta})|y \in S_h] = 0$. From Lemma 9, we know that for $1 \leq h \leq H-1$, $\boldsymbol{\epsilon}_{h,i}(\boldsymbol{\beta})$ can be treated as $c-1$ *i.i.d.* samples from $\boldsymbol{\epsilon}(\boldsymbol{\beta})|_{y\in S_h}$. According to Lemma 17 (iv),

$$\mathbb{P}\left(|\frac{1}{c-1}\sum_{i=1}^{c-1}\boldsymbol{\epsilon}_{h,i}(\boldsymbol{\beta})| > \sqrt{b}/2\right) \leq C_1\exp\left(\frac{-b(c-1)}{8C_2HK^2 + 4\sqrt{b}K}\right).$$

Similarly, we have

$$\mathbb{P}\left(|\frac{1}{c}\sum_{i=1}^{c}\boldsymbol{\epsilon}_{H,i}(\boldsymbol{\beta})| > \sqrt{b}/2\right) \leq C_1\exp\left(\frac{-bc}{8C_2HK^2 + 4\sqrt{b}K}\right).$$

Thus, if $b = O(1)$ and H is sufficiently large, we have

$$\begin{aligned}\mathbb{P}(I > \frac{b}{4}) \leq & C_1 \left((H-1) \exp\left(\frac{-b(c-1)}{8C_2 H K^2 + 4\sqrt{b}K} \right) + \exp\left(\frac{-bc}{8C_3 H K^2 + 4\sqrt{b}K} \right) \right) \\ \leq & C_1 \exp\left(-C_2 \frac{cb}{H} + C_3 \log(H) \right)\end{aligned}$$

for some positive constants C_1, C_2 and C_3.

Since $\boldsymbol{\epsilon}_i(\boldsymbol{\beta})$ are i.i.d. samples from a sub-Gaussian distribution $\boldsymbol{\epsilon}(\boldsymbol{\beta})$ with mean 0 and upper-exponentially bounded by $2K$. Lemma 19 implies that if $b = O(1)$ and H is sufficiently large, we have

$$\begin{aligned}\mathbb{P}(II > b/4) \leq & \mathbb{P}(\frac{1}{n}\sum_i \boldsymbol{\epsilon}_i(\boldsymbol{\beta})^2 > bc/4) \\ \leq & \mathbb{P}(\frac{1}{n}\sum_i \boldsymbol{\epsilon}_i(\boldsymbol{\beta})^2 - \mathbb{E}[\boldsymbol{\epsilon}(\boldsymbol{\beta})^2] > bc/4 - \mathbb{E}[\boldsymbol{\epsilon}(\boldsymbol{\beta})^2) \\ \leq & \mathbb{P}\left(\left| \frac{1}{n}\sum_i \boldsymbol{\epsilon}_i(\boldsymbol{\beta})^2 - \mathbb{E}[\boldsymbol{\epsilon}(\boldsymbol{\beta})^2] \right| \geq cb/4 - 4K^2 \right) \\ \leq & C_1 \exp\left(-C_2 \frac{\sqrt{n}(cb/4 - 4K^2)}{K^2} \right) \\ \leq & C_1 \exp\left(-C_2 \frac{cb}{H} + C_3 \log(H) \right)\end{aligned}$$

for some positive constants C_1, C_2 and C_3 if H is sufficiently large. We used in above the fact that $\mathbb{E}[\boldsymbol{\epsilon}(\boldsymbol{\beta})^2] \leq 4K^2$.

To summarize, if $b = O(1)$ and H is sufficiently large, we have

$$\mathbb{P}(var_H(\boldsymbol{x}(\boldsymbol{\beta})) > b) \leq C_1 \exp\left(-C_2 \frac{cb}{H} + C_3 \log(H) \right)$$

for some positive absolute constants C_1, C_2 and C_3.

A.2. *Proof of Lemma 1 ii)* Since $\boldsymbol{x}$ is sub-Gaussian and $\boldsymbol{\beta}$ is unit vector, we know that $var(\boldsymbol{m}(\boldsymbol{\beta})) = O(1)$. If $\boldsymbol{m}(\boldsymbol{\beta}) \neq 0$ (or equivalently $var(\boldsymbol{m}(\boldsymbol{\beta})) \neq$

0), we have

$$\begin{aligned}
&\Big|var_H(\boldsymbol{x}(\boldsymbol{\beta})) - var(\boldsymbol{m}(\boldsymbol{\beta}))\Big| \\
=&\Big|\frac{1}{H}\sum_h \overline{\boldsymbol{x}}_{h,\cdot}(\boldsymbol{\beta})^2 - var(\boldsymbol{m}(\boldsymbol{\beta}))\Big| \\
=&\Big|\frac{1}{H}\sum_h \overline{\boldsymbol{m}}_{h,\cdot}(\boldsymbol{\beta})^2 + \frac{2}{H}\sum_h \overline{\boldsymbol{m}}_{h,\cdot}(\boldsymbol{\beta})\bar{\boldsymbol{\epsilon}}_{h,\cdot}(\boldsymbol{\beta}) + \frac{1}{H}\sum_h \bar{\boldsymbol{\epsilon}}_{h,\cdot}(\boldsymbol{\beta})^2 \\
&- var(\boldsymbol{m}(\boldsymbol{\beta}))\Big| \\
\leq& A_1 + A_2 + A_3 + A_4,
\end{aligned}$$

where

$$\begin{aligned}
A_1 &= \Big|\frac{1}{H}\sum_h \mu_h(\boldsymbol{\beta})^2 - var(\boldsymbol{m}(\boldsymbol{\beta}))\Big|, \\
A_2 &= \frac{1}{H}\sum_h \Big|\overline{\boldsymbol{m}}_{h,\cdot}(\boldsymbol{\beta})^2 - \mu_h(\boldsymbol{\beta})^2\Big|, \\
A_3 &= \frac{1}{H}\sum_h \bar{\boldsymbol{\epsilon}}_{h,\cdot}(\boldsymbol{\beta})^2, \\
A_4 &= (\frac{1}{H}\sum_h \overline{\boldsymbol{m}}_{h,\cdot}(\boldsymbol{\beta})^2)^{1/2}(\frac{1}{H}\sum_h \bar{\boldsymbol{\epsilon}}_{h,\cdot}(\boldsymbol{\beta})^2)^{1/2}.
\end{aligned} \tag{14}$$

Lemma 1 ii) is a direct corollary of the following properties of A_i's.

LEMMA 2. *Let the A_i's be defined as in equation* (14). *There exist positive constants C_1, C_2 and C_3, such that for any $\nu > 1$ and H satisfying $H^\vartheta = N_1\nu$ for sufficiently large N_1, we have that each of the following events*

i) $\Theta_1 = \Big\{\ A_1 \leq \frac{1}{4\nu} var(\boldsymbol{m}(\boldsymbol{\beta}))\ \Big\}$,

ii) $\Theta_2 = \Big\{\ A_2 \leq \frac{1}{8\nu} var(\boldsymbol{m}(\boldsymbol{\beta}))\ \Big\}$,

iii) $\Theta_3 = \Big\{\ A_3 \leq \frac{1}{16\nu} var(\boldsymbol{m}(\boldsymbol{\beta}))\ \Big\}$,

iv) $\Theta_4 = \Big\{\ A_4 \leq \frac{1}{16\nu} var(\boldsymbol{m}(\boldsymbol{\beta}))\ \Big\}$,

occurs with probability at least

$$1 - C_1 \exp\left(-C_2 \frac{c\ var(\boldsymbol{m}(\boldsymbol{\beta}))}{H\nu^2} + C_3 \log(H)\right). \tag{15}$$

□

A.2.1. *Proof of Lemma 2.*

A.2.1.1. Proof of **i)** : Recall definitions of the random intervals $S_h, h = 1, 2, \cdots, H$ and random variable $\delta_h = \delta_h(\omega) = \int_{y \in S_h(\omega)} f(y)dy$. We have

$$\begin{aligned}
&\left| \frac{1}{H} \sum_h \left(\mu_h(\boldsymbol{\beta})\right)^2 - var(\boldsymbol{m}(\boldsymbol{\beta})) \right| \\
\leq &\left| var(\boldsymbol{m}(\boldsymbol{\beta})) - \sum_h \delta_h \left(\mu_h(\boldsymbol{\beta})\right)^2 \right| + \left| \frac{1}{H} \sum_h \left(\mu_h(\boldsymbol{\beta})\right)^2 - \sum_h \delta_h \left(\mu_h(\boldsymbol{\beta})\right)^2 \right| \\
\triangleq &B_1 + B_2
\end{aligned}$$

Let $\epsilon = \frac{1}{Hn_0+1}$ where $n_0 = N_2\nu$ for some sufficiently large constant N_2 and let event $E(\epsilon)$ be defined as in Lemma 11 in Section E, i.e., $E(\epsilon) = \left\{ \omega \;\middle|\; |\delta_h - \frac{1}{H}| > \epsilon, \forall h \right\}$. For any $\omega \in E(\epsilon)^c$, we have

$$\begin{aligned}
B_1 = &\sum_h \delta_h(\omega) var(\boldsymbol{m}(\boldsymbol{\beta})|y \in S_h(\omega)) \\
\leq &(\frac{1}{H} + \epsilon) \sum_h var(\boldsymbol{m}(\boldsymbol{\beta})|y \in S_h(\omega)) \qquad (16) \\
\leq &(1 + H\epsilon) \frac{\gamma_3}{H^\vartheta} var(\boldsymbol{m}(\boldsymbol{\beta})) \qquad (17) \\
\leq &\frac{2\gamma_3}{N_1\nu} var(\boldsymbol{m}(\boldsymbol{\beta})), \qquad (18)
\end{aligned}$$

where inequality (16) follows from the fact that $\delta_h(\omega) \leq \frac{1}{H} + \epsilon$, inequality (17) follows from the sliced stable condition (4) and inequality (18) follows from the requirement that $H^\vartheta > N_1\nu$, and the fact

$$\begin{aligned}
B_2 \leq &\epsilon \sum_h \left(\boldsymbol{\beta}^\tau \mu_h\right)^2 = \sum_h \frac{\epsilon}{\delta_h} \delta_h \left(\boldsymbol{\beta}^\tau \mu_h\right)^2 \\
\leq &\frac{H\epsilon}{1 - H\epsilon} \sum_h \delta_h \left(\boldsymbol{\beta}^\tau \mu_h\right)^2 \qquad (19) \\
\leq &\frac{2}{N_2\nu} \sum_h \delta_h \left(\boldsymbol{\beta}^\tau \mu_h\right)^2
\end{aligned}$$

where inequality (19) follows from the fact $\delta_h \geq \frac{1}{H} - \epsilon$.

From (17), we observe that

$$\sum_h \delta_h \left(\mu_h(\boldsymbol{\beta})\right)^2 \leq \left(1 + \frac{2\boldsymbol{\gamma}_3}{N_1\nu}\right) var(\boldsymbol{m}(\boldsymbol{\beta})). \tag{20}$$

Combining with (19), we then have

$$B_2 \leq \frac{2}{N_2\nu}\left(1 + \frac{2\boldsymbol{\gamma}_3}{N_1\nu}\right) var(\boldsymbol{m}(\boldsymbol{\beta})).$$

So when $E(\epsilon)^c$ occurs, we have

$$B_1 + B_2 \leq \left(\frac{2\boldsymbol{\gamma}_3}{N_1\nu} + \frac{2}{N_2\nu}\left(1 + \frac{2\boldsymbol{\gamma}_3}{N_1\nu}\right)\right) var(\boldsymbol{m}(\boldsymbol{\beta})).$$

Note that N_1 and N_2 can be chosen sufficiently large so that

$$B_1 + B_2 \leq \frac{4\boldsymbol{\gamma}_3}{N_1\nu} var(\boldsymbol{m}(\boldsymbol{\beta})) \leq \frac{1}{4\nu} var(\boldsymbol{m}(\boldsymbol{\beta})). \tag{21}$$

Consequently, conditioning on $E(\epsilon)^c$ where $\epsilon = \frac{1}{HN_2\nu+1}$, if we choose $H^\vartheta > N_1\nu$, then

$$\left|\frac{1}{H}\sum_h \left(\mu_h(\boldsymbol{\beta})\right)^2 - var(\boldsymbol{m}(\boldsymbol{\beta}))\right| \leq \frac{1}{4\nu} var(\boldsymbol{m}(\boldsymbol{\beta})). \tag{22}$$

Since $var(\boldsymbol{m}(\boldsymbol{\beta})) = O(1)$, $H^\vartheta > N_1\nu$ and $\epsilon = \frac{1}{HN_2\nu+1}$, the desired probability bound follows from Lemma 11, i.e.,

$$\begin{aligned}\mathbb{P}(E(\epsilon)) \leq & C_1 \exp\left(-\frac{Hc+1}{32(Hn_0+1)^2} + \log(H^2\sqrt{Hc+1})\right) \\ \leq & C_1 \exp\left(-C_2\frac{c\ var(\boldsymbol{m}(\boldsymbol{\beta}))}{H\nu^2} + C_3\log(H)\right).\end{aligned}$$

for some positive constants C_1, C_2 and C_3. □

Remark 5. From (22), conditioning on $E(\epsilon)^c$, we obtain the following two inequalities

$$\frac{1}{H}\sum_h \left(\mu_h(\boldsymbol{\beta})\right)^2 \leq \left(1 + \frac{4\boldsymbol{\gamma}_3}{H^\vartheta}\right) var\left(\boldsymbol{m}(\boldsymbol{\beta})\right) \tag{23}$$

and

$$\frac{1}{H}\sum_h |\mu_h(\boldsymbol{\beta})| \leq \left(\left(1 + \frac{4\boldsymbol{\gamma}_3}{H^\vartheta}\right) var\left(\boldsymbol{m}(\boldsymbol{\beta})\right)\right)^{1/2}. \tag{24}$$

In particular, $\frac{1}{H}\sum_h \left(\mu_h(\boldsymbol{\beta})\right)^2$ and $\frac{1}{H}\sum_h |\mu_h(\boldsymbol{\beta})|$ are bounded by $O_P(1)$.

A.2.1.2. Proof of **ii)** : Denote $\frac{1}{c-1}\sum_{i=1}^{c-1}\boldsymbol{m}_{h,i}(\boldsymbol{\beta})$ by $\overline{\boldsymbol{m}}'_h(\boldsymbol{\beta})$ and $\overline{\boldsymbol{m}}_{H,\cdot}(\boldsymbol{\beta})$ by $\overline{\boldsymbol{m}}'_H(\boldsymbol{\beta})$, we have

$$\begin{aligned}A_2 \leq & \frac{1}{H}\sum_{h=1}^{H}\left|\overline{\boldsymbol{m}}'_h(\boldsymbol{\beta})^2-\mu_h(\boldsymbol{\beta})^2\right|+\frac{1}{Hc^2}\sum_{h=1}^{H}\boldsymbol{m}_{h,c}(\boldsymbol{\beta})^2\\ &+\frac{2(c-1)}{c}\left(\frac{1}{H}\sum_{h=1}^{H}\overline{\boldsymbol{m}}'_h(\boldsymbol{\beta})^2\right)^{1/2}\left(\frac{1}{Hc^2}\sum_{h=1}^{H}\boldsymbol{m}_{h,c}(\boldsymbol{\beta})^2\right)^{1/2}+\frac{2}{Hc}\sum_{h=1}^{H}\mu_h(\boldsymbol{\beta})^2\\ \triangleq & \ I+II+III+IV\end{aligned}$$

Before we start proving this part, we need to introduce two events and bound their probabilities. First, let

$$E_1(N_3,\nu)=\left\{\ \eta(\boldsymbol{\beta})>\frac{1}{N_3\nu}\sqrt{var(\boldsymbol{m}(\boldsymbol{\beta}))}\ \right\}. \tag{25}$$

where $\eta(\boldsymbol{\beta})=\max_{1\leq h\leq H}\left\{\left|\overline{\boldsymbol{m}}'_h(\boldsymbol{\beta})-\mu_h(\boldsymbol{\beta})\right|\right\}$. According to Lemma 17 (i), (iv) and Bonferroni's inequality, we have

$$\begin{aligned}\mathbb{P}\left(E_1(N_3,\nu)\right)\leq & 2H\exp\left(\frac{1}{(N_3\nu)^2}\frac{-(c-1)var(\boldsymbol{m}(\boldsymbol{\beta}))}{2CHK^2+\frac{2}{N_3\nu}\sqrt{var(m(\boldsymbol{\beta}))}K}\right) &(26)\\ \leq & C_1\exp\left(-C_2\frac{c\ var(\boldsymbol{m}(\boldsymbol{\beta}))}{H\nu^2}+C_3\log(H)\right) &(27)\end{aligned}$$

for some positive constants C_1, C_2 and C_3. Second, let

$$E_2(N_4,\nu)\triangleq\left\{\ II>\frac{1}{N_4\nu}var(\boldsymbol{m}(\boldsymbol{\beta}))\ \right\},$$

then

$$\begin{aligned}\mathbb{P}\left(E(N_4,\nu)\right)\leq & \mathbb{P}\left(\frac{1}{nc}\sum_i m_i^2>\frac{var(\boldsymbol{m}(\boldsymbol{\beta}))}{N_4\nu}\right)\\ \leq & C_1\exp\left(-C_2\sqrt{n}(c\ \frac{var(\boldsymbol{m}(\boldsymbol{\beta}))}{\nu})-K^2)\right)\\ \leq & C_1\exp\left(-C_2\frac{c\ var(\boldsymbol{m}(\boldsymbol{\beta}))}{H\nu}+C_3\log(H)\right)\end{aligned}$$

for some positive constant C_1, C_2 and C_3. It is easily to see $E(N_4,\nu)\subset E(N_4,\nu^2)$.

For I. Conditioning on the event $E(\epsilon)^c \cap E_1(N_3,\nu)^c$, combining with (24), we have

$$\begin{aligned} I \leq & \frac{1}{H}\sum_h \eta(\boldsymbol{\beta})(\eta(\boldsymbol{\beta})+2|\mu_h(\boldsymbol{\beta})|) \leq \eta(\boldsymbol{\beta})^2 + \frac{2\eta(\boldsymbol{\beta})}{H}\sum_h |\mu_h(\boldsymbol{\beta})| \\ \leq & \left(\left(\frac{1}{N_3\nu}\right)^2 + \frac{2}{N_3\nu}\left(1+\frac{4\gamma_3}{H^\vartheta}\right)^{1/2}\right) var(\boldsymbol{m}(\boldsymbol{\beta})) \\ \leq & \frac{1}{32\nu} var(\boldsymbol{m}(\boldsymbol{\beta})) \end{aligned}$$

if N_3 is sufficiently large .

REMARK 6. From above, conditioning on the event $E(\epsilon)^c \cap E_1(N_3,\nu)^c$, we have

$$\frac{1}{H}\sum_{h=1}^{H} \overline{\boldsymbol{m}}'(\boldsymbol{\beta})^2 \leq \frac{1+32\nu}{32\nu} var(\boldsymbol{m}(\boldsymbol{\beta})). \tag{28}$$

For II. Conditioning on $E_2(N_4,\nu)^c$, we have $II \leq \frac{var(\boldsymbol{m}(y))}{N_4\nu}$.

For III. When the event $E(\epsilon)^c \cap E_1(N_3,\nu)^c \cap E_2(N_4,\nu^2)^c$ occurs, according to equation (28),

$$III \leq \frac{2(c-1)}{c}\sqrt{\frac{1+32\nu}{32\nu}}\frac{1}{\sqrt{N_4}\nu} var(\boldsymbol{m}(\boldsymbol{\beta})) < \frac{1}{16\nu} var(\boldsymbol{m}(\boldsymbol{\beta})).$$

if N_4 is sufficiently large.

For VI. When the event $E(\epsilon)^c \cap E_1\,(N_3,\nu)^c \cap E_2(N_4,\nu)^c$ occurs, from (22), we know

$$VI = \frac{2}{Hc}\sum_h \mu_h(\boldsymbol{\beta})^2 \leq \frac{9}{4c} var(\boldsymbol{m}(\boldsymbol{\beta})) < \frac{1}{16\nu} var(\boldsymbol{m}(\boldsymbol{\beta})).$$

To summarize, we know that there exist positive constant C_1, C_2, C_3 and C_4 such that

$$A_2 \leq I + II + III + VI \leq \frac{1}{8\nu} var(\boldsymbol{m}(\boldsymbol{\beta}))$$

holds on the event $E(\epsilon)^c \cap E_1(N_3,\nu)^c \cap E_2(N_4,\nu^2)^c$ which is with probability at least

$$1 - C_1 \exp\left(-C_2 \frac{c\ var(\boldsymbol{m}(\boldsymbol{\beta}))}{H\nu^2} + C_3 \log(H)\right)$$

for some positive constants C_1,C_2 and C_3.

A.2.1.3. Proof of **iii**) : Similar to the proof of Lemma 1 (i) we have

$$\mathbb{P}(A_3 > b) \leq C_1 H \exp\left(\frac{-(c-1)b}{8C_2 H K_1^2 + 4\sqrt{b}K_2}\right)$$

for some positive constants C_1, C_2 and C_3. In particular , if we take $b = \frac{1}{16\nu} var(\boldsymbol{m}(\boldsymbol{\beta}))$, we know that

$$A_3 \leq \frac{1}{16\nu} var(\boldsymbol{m}(\boldsymbol{\beta}))$$

with probability at least

$$1 - C_1 \exp\left(-C_2 \frac{c\ var(\boldsymbol{m}(\boldsymbol{\beta}))}{H\nu^2} + C_3 \log(H)\right)$$

for some positive constant C_1, C_2 and C_3 .

A.2.1.4. Proof of **iv**) : Let

$$D_1 \triangleq \frac{1}{H}\sum_h \overline{\boldsymbol{m}}_{h,\cdot}(\boldsymbol{\beta})^2, \quad D_2 \triangleq A_3 = \frac{1}{H}\sum_h \bar{\boldsymbol{\epsilon}}_{h,\cdot}(\boldsymbol{\beta})^2$$

Consequently,

$$\begin{aligned} &\mathbb{P}\left(D_1^{1/2} D_2^{1/2} > \frac{1}{16\nu} var(\boldsymbol{m}(\boldsymbol{\beta}))\right) \\ \leq &\mathbb{P}\left(|D_1| > \frac{2\nu+1}{2\nu} var(\boldsymbol{m}(\boldsymbol{\beta}))\right) + \mathbb{P}\left(D_2 > \frac{var(\boldsymbol{m}(\boldsymbol{\beta}))}{(2\nu+1)16\nu}\right) \end{aligned} \tag{29}$$

Note that

$$|D_1 - var(\boldsymbol{m}(\boldsymbol{\beta}))| \leq A_2 + A_1$$

According to (i) and (ii), the right hand side of (29) is bounded by

$$C_1 \exp\left(-C_2 \frac{c\ var(\boldsymbol{m}(\boldsymbol{\beta})}{H\nu^2} + C_3 \log(H)\right)$$

for some positive constants C_1, C_2 and C_3.

□

B. Proofs of theorems in section 3.

B.1. *Proof of Theorem 1.* Let $\mathcal{S}$ be the central subspace of dimension $d \ll p$, i.e., $y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{P}_{\mathcal{S}}\boldsymbol{x}$ and $dim(\mathcal{S}) = d$. We have the decomposition

$$\begin{aligned} \boldsymbol{x} &= \boldsymbol{P}_{\mathcal{S}}\boldsymbol{x} + \boldsymbol{P}_{\mathcal{S}^\perp}\boldsymbol{x} \triangleq \boldsymbol{z} + \boldsymbol{w} \\ &= \mathbb{E}[\boldsymbol{z}|y] + \boldsymbol{z} - \mathbb{E}[\boldsymbol{z}|y] + \boldsymbol{w} \triangleq \boldsymbol{m} + \boldsymbol{v} + \boldsymbol{w} \end{aligned} \tag{30}$$

where $\boldsymbol{z} = \boldsymbol{P}_{\mathcal{S}}\boldsymbol{x}$, $\boldsymbol{m} = \mathbb{E}[\boldsymbol{z}|y]$, $\boldsymbol{v} = \boldsymbol{z} - \mathbb{E}[\boldsymbol{z}|y]$ and $\boldsymbol{w} = \boldsymbol{P}_{\mathcal{S}^\perp}\boldsymbol{x}$. Note that $\boldsymbol{m}$ lies in the central curve, $\boldsymbol{v}$ lies in the central space and $\boldsymbol{w}$ lies in the space perpendicular to $\mathcal{S}$. We introduce

$$\boldsymbol{m}_{h,j},\ \overline{\boldsymbol{m}}_{h,\cdot},\ \overline{\overline{\boldsymbol{m}}}, \quad \boldsymbol{z}_{h,j},\ \overline{\boldsymbol{z}}_{h,\cdot},\ \overline{\overline{\boldsymbol{z}}}, \text{ and } \boldsymbol{w}_{h,j},\ \overline{\boldsymbol{w}}_{h,\cdot},\ \overline{\overline{\boldsymbol{w}}} \tag{31}$$

similar to the definition of $\boldsymbol{x}_{h,j}$, $\overline{\boldsymbol{x}}_{h,\cdot}$ and $\overline{\overline{\boldsymbol{x}}}$. Consequently, we can define $\widehat{\boldsymbol{\Lambda}}_{\boldsymbol{z}}$ and have the following decomposition

$$\widehat{\boldsymbol{\Lambda}}_H \equiv \frac{1}{H}\sum_h \overline{\boldsymbol{x}}_{h,\cdot}\overline{\boldsymbol{x}}_{h,\cdot}^\tau = \widehat{\Lambda}_{\boldsymbol{z}} + \mathcal{Z}\mathcal{W}^\tau + \mathcal{W}\mathcal{Z}^\tau + \mathcal{W}\mathcal{W}^\tau, \tag{32}$$

where

$$\mathcal{Z} = \frac{1}{\sqrt{H}}\left(\overline{\boldsymbol{z}}_{1,\cdot}, ..., \overline{\boldsymbol{z}}_{H,\cdot}\right) \text{ and } \mathcal{W} = \frac{1}{\sqrt{H}}\left(\overline{\boldsymbol{w}}_{1,\cdot}, ..., \overline{\boldsymbol{w}}_{H,\cdot}\right).$$

We need to bound $\|\widehat{\boldsymbol{\Lambda}}_{\boldsymbol{z}} - \boldsymbol{\Lambda}_{\boldsymbol{p}}\|_2$ and $\|\mathcal{W}\mathcal{W}^\tau\|_2$.

Lemma 3.

$$\|\mathcal{W}\mathcal{W}^\tau\|_2 \leq O_P(\frac{H^2 p}{n}) \tag{33}$$

Proof. From Lemma 1, for any unit vector $\boldsymbol{\beta} \perp col(\boldsymbol{\Lambda})$, i.e. $var(\boldsymbol{m}(\boldsymbol{\beta})) = 0$, we have

$$\mathbb{P}(\boldsymbol{\beta}^\tau\mathcal{W}\mathcal{W}^\tau\boldsymbol{\beta} > C\frac{H^2 p}{n}) \leq C_1 \exp\left(-C_2 p + \log(H)\right). \tag{34}$$

for some positive constants C_1 and C_2. Then the ε-net argument (see e.g., Vershynin [2010]) implies that $\|\mathcal{W}\mathcal{W}^\tau\| \leq O_P(\frac{H^2 p}{n})$ □

Lemma 4.

$$\|\widehat{\boldsymbol{\Lambda}}_{\boldsymbol{z}} - \boldsymbol{\Lambda}_p\| \leq O_P\left(\frac{1}{H^\vartheta}\right). \tag{35}$$

As a direct corollary, we have $\|\widehat{\boldsymbol{\Lambda}}_{\boldsymbol{z}}\| \leq O_P(1)$.

PROOF. From Lemma 1, we have

$$\mathbb{P}\left(\left|\boldsymbol{\beta}^\tau(\widehat{\boldsymbol{\Lambda}}_{\boldsymbol{z}}-\boldsymbol{\Lambda})\boldsymbol{\beta}\right|>\frac{C}{H^\vartheta}\|\boldsymbol{\Lambda}\|_2\right)\le C_1\exp\left(-C_2\frac{c\ var(\boldsymbol{m}(\boldsymbol{\beta}))}{H^{1+2\vartheta}}+C_3\log(H)\right).$$

Note that we only need to verify it for $\boldsymbol{\beta}\in col(\boldsymbol{\Lambda}_p)$, which is a d-dimensional space. Then the ε-net argument implies that $\|\widehat{\boldsymbol{\Lambda}}_{\boldsymbol{z}}-\boldsymbol{\Lambda}_p\|_2\le O_P\left(\frac{1}{H^\vartheta}\right)$. □

Theorem 1 follows from Lemma 4 and Lemma 3. In fact,

$$\begin{aligned}\|\widehat{\boldsymbol{\Lambda}}_H-\boldsymbol{\Lambda}_p\|&\le\|\widehat{\boldsymbol{\Lambda}}_{\boldsymbol{z}}-\boldsymbol{\Lambda}_p\|+\|\mathcal{Z}\mathcal{W}^\tau+\mathcal{W}\mathcal{Z}^\tau\|_2+\|\mathcal{W}\mathcal{W}^\tau\|_2\\&\le O_P\left(\frac{1}{H^\vartheta}+\sqrt{\frac{H^2p}{n}}+\frac{H^2p}{n}\right).\end{aligned}$$

□

B.2. *Proof of Theorem 2.* Theorem 2 is a direct corollary of Theorem 1 and Lemma 13. In fact, we have:

$$\begin{aligned}&\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X}}^{-1}\widehat{\boldsymbol{\Lambda}}_H-\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Lambda}_p\|_2\\&\le\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X}}^{-1}-\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\|_2\|\widehat{\boldsymbol{\Lambda}}_H\|_2+\|\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\|_2\|\widehat{\boldsymbol{\Lambda}}_H-\boldsymbol{\Lambda}_p\|_2,\end{aligned}$$

which $\to 0$ if $\rho=\lim_{n\to\infty}\frac{p}{n}=0$. □

B.3. *Proof of Theorem 3.*

(i) The proof for part (i) is similar to the proof of Theorem 1 and the standard Gaussian assumption on $\boldsymbol{x}$ simplifies the argument and improves the results. Since $\boldsymbol{w}=\boldsymbol{P}_{\mathcal{S}^\perp}\boldsymbol{x}$ is normal and independent of y, there exists a normal random variable $\boldsymbol{\epsilon}\sim N(\mathbf{0},\mathbf{I})$ such that $\boldsymbol{w}=\boldsymbol{\Sigma}^{1/2}\boldsymbol{\epsilon}$ where $\boldsymbol{\Sigma}=cov(\boldsymbol{w})$. Using the decomposition (32), we may write

$$\mathcal{W}=\frac{1}{\sqrt{Hc}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{E}_{p\times H}\tag{36}$$

where $\boldsymbol{E}_{p,H}$ is a $p\times H$ matrix with $i.i.d.$ standard normal entries. Corollary 4 implies that

$$\|\mathcal{W}\mathcal{W}^\tau\|_2\le C\left(\sqrt{\frac{p}{n}}+\sqrt{\frac{H}{n}}\right)^2\le O_P\left(\frac{p}{n}\right).$$

Lemma 4 implies

$$\|\widehat{\mathbf{\Lambda}}_{\boldsymbol{z}}\|_2 \leq \|\mathbf{\Lambda}_p\|_2 + O_P\left(\frac{1}{H^{\vartheta}}\right).$$

By the Cauchy inequality, we have

$$\|\mathcal{C}_1\|_2^2 \leq \|\widehat{\mathbf{\Lambda}}_{\boldsymbol{z}}\|_2 \|\mathcal{W}\mathcal{W}^{\tau}\|_2 \leq O_P\left(\frac{p}{n}\right).$$

Thus,

$$\|\widehat{\mathbf{\Lambda}}_H - \mathbf{\Lambda}_p\|_2 \leq O_P(\frac{1}{H^{\vartheta}} + \frac{p}{n} + \sqrt{\frac{p}{n}}).$$

In particular, if $H, n \to \infty$ and $\rho = \lim \frac{p}{n} \in (0, \infty)$, we know that $\|\widehat{\mathbf{\Lambda}}_H - \mathbf{\Lambda}_p\|_2$ is dominated by $\rho \vee \sqrt{\rho}$ as a function of ρ. □

(ii) The proof for part (ii) is similar to the proof of Theorem 2 in Johnstone and Lu [2009] but is technically more challenging. Let $D = \mathcal{Z}\mathcal{Z}^{\tau} + \mathcal{W}\mathcal{W}^{\tau}$ and $B = \mathcal{Z}\mathcal{W}^{\tau} + \mathcal{W}\mathcal{Z}^{\tau}$, then

$$\widehat{\mathbf{\Lambda}}_H = D + B.$$

Since we are working on single index model with $\boldsymbol{x}$ is standard normal, $\boldsymbol{z} = P_{\boldsymbol{\beta}}\boldsymbol{x} = \boldsymbol{\beta} z(y)$ for some scalar function $z(y)$ and $\boldsymbol{w} = P_{\boldsymbol{\beta}^{\perp}}\boldsymbol{x}$ are independent normal random variables. Let $\mathbf{\Sigma} = var(\boldsymbol{w})$, then we can write

$$\mathcal{W} = \frac{1}{\sqrt{Hc}}\mathbf{\Sigma}^{1/2}\mathbf{E}$$

where $\mathbf{E}$ is a $p \times H$ matrix with *i.i.d.* standard normal entries.

Since $\boldsymbol{z} = \boldsymbol{\beta} z(y)$, we have $\mathcal{Z} = \frac{1}{\sqrt{H}}\boldsymbol{\beta}(\overline{z}_{1,\cdot}, \overline{z}_{2,\cdot}, ..., \overline{z}_{H,\cdot})$. To ease notation, let $\boldsymbol{\theta}^{\tau} = (\overline{z}_{1,\cdot}, \overline{z}_{2,\cdot}, ..., \overline{z}_{H,\cdot})$, then

$$\begin{aligned} D &= \frac{1}{H}\|\boldsymbol{\theta}\|^2 \boldsymbol{\beta}\boldsymbol{\beta}^{\tau} + \frac{1}{n}\mathbf{\Sigma}^{1/2}\boldsymbol{E}\boldsymbol{E}^{\tau}\mathbf{\Sigma}^{1/2} \\ B &= \boldsymbol{\beta}\boldsymbol{u}^{\tau} + \boldsymbol{u}\boldsymbol{\beta}^{\tau} \text{ where } \quad \boldsymbol{u} = \frac{1}{H\sqrt{c}}\mathbf{\Sigma}^{1/2}\boldsymbol{E}\boldsymbol{\theta}. \end{aligned} \tag{37}$$

Let $0 < \alpha < arctan(\frac{1}{16})$ and

$$N_{\alpha} = \left\{ \boldsymbol{x} \in \mathbb{R}^p : \angle(\boldsymbol{x}, \boldsymbol{\beta}) \leq \alpha \text{ and } \|x\| = 1 \right\} \tag{38}$$

be the set of unit vectors making angle at most α where $\angle(\boldsymbol{x}, \boldsymbol{y})$ is the angle between the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. In order to proceed, we need the following lemma.

LEMMA 5. *Let $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_-$ be the principal eigenvector of $S_+ \triangleq D + B$ and $S_- \triangleq D - B$, respectively. There exists a positive constant $\omega(\alpha)$ such that for any $\widehat{\boldsymbol{\beta}} \in N_\alpha$, i.e., $\angle\left(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}\right) \leq \alpha$, we have*

$$\angle\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}_-\right) \geq \frac{1}{3}\omega(\alpha) \tag{39}$$

with probability converging to one as $n \to \infty$.

PROOF. The proof is presented in Lin et al. [2015].

Note that S_+ and S_- have the same distribution (viewed as functions of random terms $\boldsymbol{E}$ and θ):

$$S_-(\boldsymbol{E}, \theta) = S_+(-\boldsymbol{E}, \theta).$$

Let $\mathcal{A}_\alpha$ denote the event $\left\{\angle\left(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}\right) \leq \alpha\right\} \cup \left\{\angle\left(\widehat{\boldsymbol{\beta}}_-, \boldsymbol{\beta}\right) \leq \alpha\right\}$, then

$$\begin{aligned}
\mathbb{E}[\angle\left(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}\right)] &\geq \mathbb{E}[\angle\left(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}\right), \mathcal{A}_\alpha^c] + \mathbb{E}[\angle\left(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}\right), \mathcal{A}_\alpha] \\
&\geq \mathbb{E}[\angle\left(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}\right), \mathcal{A}_\alpha^c] + \frac{1}{2}\mathbb{E}[\angle\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}_-\right), \mathcal{A}_\alpha] \\
&\geq \min\{\alpha, \frac{\omega(\alpha)}{6}\} > 0.
\end{aligned}$$

□

SUPPLEMENTARY MATERIAL

Supplement to " On the consistency and sparsity for sliced inverse regression for high dimensions" (http://www.e-publications.org/ims/support/dowload/imsart-ims.zip). In the supplement, we prove the rest results stated in the paper.

References

P. J. Bickel and E. Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.

T. T. Cai, C. Zhang, and H. Zhou. Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144, 2010.

E. Candes and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351, 2007.

R. D. Cook. Graphics for regressions with a binary response. *Journal of the American Statistical Association*, 91(435):983–992, 1996.

R. D. Cook, L. Forzani, and A. J. Rothman. Estimating sufficient reductions of the predictors in abundant high-dimensional regressions. *The Annals of Statistics*, 40(1): 353–384, 2012.

H. Cui, R. Li, and W. Zhong. Model-free feature screening for ultrahigh dimensional discriminant analysis. *Journal of the American Statistical Association*, 110(510):630–641, 2015.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B*, 70(5):849–911, 2008.

T. Hsing and R. J. Carroll. An asymptotic theory for sliced inverse regression. *The Annals of Statistics*, 20(2):1040–1061, 1992.

B. Jiang and J. S. Liu. Variable selection for general index models via sliced inverse regression. *The Annals of Statistics*, 42(5):1751–1786, 2014.

I. M. Johnstone and A. Y. Lu. Sparse principal components analysis. 2004.

I. M. Johnstone and A. Y. Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486), 2009.

K. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

L. Li. Sparse sufficient dimension reduction. *Biometrika*, 94(3):603–613, 2007.

L. Li and C. J. Nachtsheim. Sparse sliced inverse regression. *Technometrics*, 48(4), 2006.

Q. Lin, Z. Zhao, and J. S. Liu. Supplement to "on consistency and sparsity for sliced inverse regression in high dimensions". 2015.

X. Luo, L. A. Stefanski, and D. D. Boos. Tuning variable selection procedures by adding noise. *Technometrics*, 48(2):165–175, 2006.

M. Neykov, Q. Lin, and J. S. Liu. Signed support recovery for single index models in high-dimensions. *arXiv preprint arXiv:1511.02270*, 2015.

M. Rudelson and R. Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electron. Commun. Probab*, 18(82):1–9, 2013.

G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Y. Wu, D. D. Boos, and L. A. Stefanski. Controlling variable selection by the addition of pseudovariables. *Journal of the American Statistical Association*, 102(477):235–243, 2007.

Z. Yu, L. Zhu, H. Peng, and L. Zhu. Dimension reduction and predictor selection in semiparametric models. *Biometrika*, page ast005, 2013.

Z. Yu, Y. Dong, and L. X. Zhu. Trace pursuit: A general framework for model-free variable selection. *Journal of the American Statistical Association*, 2016. To appear.

W. Zhong, T. Zhang, Y. Zhu, and J. S. Liu. Correlation pursuit: forward stepwise variable selection for index models. *Journal of the Royal Statistical Society: Series B*, 74(5):849–870, 2012.

L. Zhu and K. Fang. Asymptotics for kernel estimate of sliced inverse regression. *The Annals of Statistics*, 24(3):1053–1068, 1996.

L. Zhu, L. Li, R. Li, and L. Zhu. Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 106(496), 2011.

L. X. Zhu and K. W. Ng. Asymptotics of sliced inverse regression. *Statistica Sinica*, 5(2): 727–736, 1995.

L. X. Zhu, B. Miao, and H. Peng. On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association*, 101(474), 2006.

Li-Ping Zhu, Lexin Li, Runze Li, and Li-Xing Zhu. Model-free feature screening for

ultrahigh-dimensional data. *Journal of the American Statistical Association*, 2012.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320, 2005.

Submitted to the Annals of Statistics
arXiv: arXiv:0000.0000

SUPPLEMENT TO "ON CONSISTENCY AND SPARSITY FOR SLICED INVERSE REGRESSION IN HIGH DIMENSIONS"

By Qian Lin[§] Zhigen Zhao[¶] and Jun S. Liu[§]

Harvard University[§]
Temple University[¶]

C. Proof of Lemma 5. We need the following lemmas.

Lemma 6. *Recall that* $\boldsymbol{u} = \frac{1}{H\sqrt{c}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{E}\boldsymbol{\theta}$ *defined as in* (37)*, then there exist positive constants* C_1 *and* C_2 *such that*

$$0 < C_1 \leq \|\boldsymbol{u}\|_2 \leq C_2$$

with probability converging to one as $n \to \infty$*.*

Lemma 7. *Assuming conditions in Theorem 3, let* B *and* N_α *be defined as in* (37) *and*(38) *respectively where* $0 < \alpha < arctan(\frac{1}{16})$ *.*

i) There exists positive constant C_1 *such that for any* $\boldsymbol{x} \in N_\alpha$*, we have* $\|B\boldsymbol{x}\| \geq C_1$ *with probability converging to one as* $n \to \infty$*;*

ii) For any $\boldsymbol{x} \in N_\alpha$*, we have* $\left|\cos\angle(\boldsymbol{x}, B\boldsymbol{x})\right| \leq 4\alpha$ *with probability converging to one as* $n \to \infty$*.*

The following lemma is borrowed from Johnstone and Lu [2004].

Lemma 8. *Let* $\boldsymbol{\xi}$ *be a principal eigenvector of a non-zero symmetric matrix* M*. For any* $\boldsymbol{\eta} \neq 0$*,*

$$\angle(\boldsymbol{\eta}, M\boldsymbol{\eta}) \leq 3\angle(\boldsymbol{\eta}, \boldsymbol{\xi}).$$

The proof of Lemma 5 is made plausible by reference to the Figure C1.

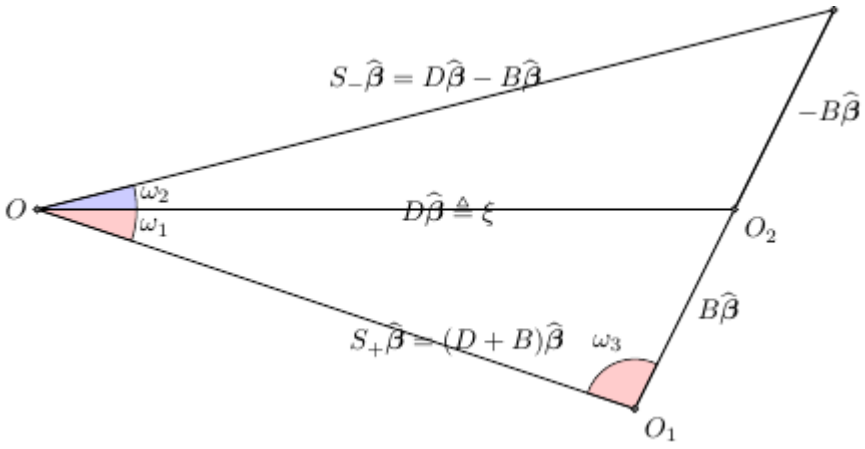

Fig C1: An illustrated graph

Since

$$\sin\left(\angle\left(\widehat{\boldsymbol{\beta}}, S_-\widehat{\boldsymbol{\beta}}\right)\right) = \sin(\omega_1+\omega_2) = \sin(\pi-\omega_1-\omega_2) \tag{C.1}$$

$$\geq \min\left\{\ \sin(\omega_1), \sin(\omega_3)\ \right\}, \tag{C.2}$$

we only need to prove that there exists a positive small constant $\omega(\alpha)$ ($< \frac{\pi}{2}$) such that $\sin(\omega_1), \sin(\omega_2) \geq \sin(\omega(\alpha))$. In fact, if such $\omega(\alpha)$ exists, we may choose $\boldsymbol{M} = S_-, \xi = \widehat{\boldsymbol{\beta}}_-$ in Lemma 8 and get

$$\angle\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}_-\right) \geq \frac{1}{3}\angle\left(\widehat{\boldsymbol{\beta}}, S_-\widehat{\boldsymbol{\beta}}\right) \geq \frac{1}{3}\omega(\alpha).$$

For ω_3. From Lemma 7 ii), $\left|\cos\angle(\widehat{\boldsymbol{\beta}}, B\widehat{\boldsymbol{\beta}})\right| \leq 4\alpha$, we know that there exists positive constants $\delta(\alpha)(< \frac{\pi}{2})$ such that $\sin\omega_3 \geq sin\,(\delta(\alpha))$.

For ω_1. Applying the law of sines to the triangle $\triangle(O, O_1, O_2)$, we have

$$\frac{\sin\omega_1}{\|B\widehat{\boldsymbol{\beta}}\|} = \frac{\sin\omega_3}{\|D\widehat{\boldsymbol{\beta}}\|}\left(= \frac{\sin\angle\left(B\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}\right)}{\|D\widehat{\boldsymbol{\beta}}\|}\right). \tag{C.3}$$

Note that from Lemma 7 i), there exists a constant $C_1 > 0$ such that $\|B\widehat{\boldsymbol{\beta}}\| > C_1$ and

$$\|D\widehat{\boldsymbol{\beta}}\| \leq \|D\| \leq \frac{1}{H}\|\theta\|^2 + \|\frac{1}{n}EE^\tau\|,$$

is bounded by an absolute constant C given $\lim_{n\to\infty}\frac{p}{n}=\rho\neq 0$ and sliced stable condition. Then (C.3) implies

$$\sin\omega_1=\frac{\|B\widehat{\boldsymbol{\beta}}\|\sin\angle(B\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\beta}})}{\|D\widehat{\boldsymbol{\beta}}\|}\geq\frac{C_1\sin\delta(\alpha)}{C}\geq\sin\omega'>0$$

where ω' ($<\frac{\pi}{2}$)is an small angle such that the last inequality holds. In particular, we have $\omega_1\geq\omega'$. Hence

$$\angle(\widehat{\boldsymbol{\beta}},S_-\widehat{\boldsymbol{\beta}})\geq\omega'\wedge\delta(\alpha)\triangleq\omega(\alpha)$$

□

C.1. *Proof of Lemma 6* *Proof* : In fact, let $\boldsymbol{T}$ be an orthogonal matrix such that $\boldsymbol{T}\boldsymbol{\beta}=(1,0\cdots,0)^\tau$ and $\boldsymbol{M}=\boldsymbol{T}\boldsymbol{\beta}\boldsymbol{\beta}^\tau\boldsymbol{T}^\tau$, then

$$\begin{aligned}cH^2\boldsymbol{u}^\tau\boldsymbol{u}&=\boldsymbol{\theta}^\tau\boldsymbol{E}^\tau\boldsymbol{\Sigma}E\boldsymbol{\theta}\\&=\boldsymbol{\theta}^\tau\boldsymbol{E}^\tau\boldsymbol{E}\boldsymbol{\theta}-\boldsymbol{\theta}^\tau\boldsymbol{E}^\tau\boldsymbol{\beta}\boldsymbol{\beta}^\tau\boldsymbol{E}\boldsymbol{\theta}\\&=\boldsymbol{\theta}^\tau\boldsymbol{E}^\tau\boldsymbol{T}^\tau\boldsymbol{T}\boldsymbol{E}\boldsymbol{\theta}-\boldsymbol{\theta}^\tau\boldsymbol{E}^\tau\boldsymbol{T}^\tau(\boldsymbol{T}\boldsymbol{\beta})\boldsymbol{\beta}^\tau\boldsymbol{T}^\tau\boldsymbol{T}\boldsymbol{E}\boldsymbol{\theta}\\&\overset{d.}{=}\boldsymbol{\theta}^\tau\boldsymbol{E}^\tau\boldsymbol{E}\boldsymbol{\theta}-\boldsymbol{\theta}^\tau\boldsymbol{E}^\tau\boldsymbol{M}\boldsymbol{E}\boldsymbol{\theta}\\&\overset{d.}{=}\frac{p-1}{p}\boldsymbol{\theta}^\tau\boldsymbol{E}^\tau\boldsymbol{E}\boldsymbol{\theta},\end{aligned}$$

where $\overset{d.}{=}$ means equal in distribution. Note that $\boldsymbol{E}^\tau\boldsymbol{E}$ is full rank $H\times H$ matrix, combining with Lemma 14 , we know that

$$C_1(1-\sqrt{\frac{H}{p}})^2\leq\lambda_{min}(\frac{1}{p}\boldsymbol{E}^\tau\boldsymbol{E})\leq\lambda_{max}(\frac{1}{p}\boldsymbol{E}^\tau\boldsymbol{E})\leq C_2(1+\sqrt{\frac{H}{p}})^2$$

for some positive constants C_1 and C_2 with probability at least $1-2\exp(-p/8)$. Note that $\lim\frac{p}{n}=\rho>0$ as $n\to\infty$ and $n=Hc$, we know there exists positive constants C_1 and C_2 such that

$$C_1\frac{1}{H}\|\boldsymbol{\theta}\|^2\leq\|\boldsymbol{u}\|^2\leq C_2\frac{1}{H}\|\boldsymbol{\theta}\|^2$$

with probability at least $1-2\exp(-p/8)$.

On the other hand, the sliced stable condition implies that $\lim\frac{1}{H}\|\boldsymbol{\theta}\|^2$ exists ($\neq 0$), so $\|\boldsymbol{u}\|^2$ is bounded away from 0 and ∞ with probability 1 as $n\to\infty$. □

C.2. *Proof of Lemma 7* *Proof*: For i), $\forall \boldsymbol{x} \in N_\alpha$, let

$$\boldsymbol{x} = \cos(\delta)\boldsymbol{\beta} + \sin(\delta)\boldsymbol{\eta} \text{ where } \boldsymbol{\eta} \perp \boldsymbol{\beta}, \|\boldsymbol{\eta}\| = 1, \delta \leq \alpha.$$

Since $B\boldsymbol{x} = \cos(\delta)\boldsymbol{u} + (\boldsymbol{u}^\tau\boldsymbol{\eta})\sin(\delta)\boldsymbol{\beta}$, we have:

$$\begin{aligned}\|B\boldsymbol{x}\| &\geq \cos(\delta)\|\boldsymbol{u}\| - \sin(\delta)\|\boldsymbol{u}\| \geq \frac{1}{2}\cos(\delta)\|\boldsymbol{u}\| \\ &\geq \frac{1}{2}\cos(\alpha)\|\boldsymbol{u}\| > \frac{C_1}{4} > 0\end{aligned}$$

for some positive constant C_1.

For ii), since

$$\boldsymbol{x}^\tau B\boldsymbol{x} = 2(\boldsymbol{u}^\tau\boldsymbol{\eta})\cos(\delta)\sin(\delta)$$

we have that uniformly over N_δ,

$$|\boldsymbol{x}^\tau B\boldsymbol{x}| \leq |\boldsymbol{u}^\tau\boldsymbol{\eta}|\sin(2\delta)$$

which in turn implies:

$$\left|\cos(\angle(B\boldsymbol{x}, \boldsymbol{x}))\right| = \frac{|\boldsymbol{x}^\tau B\boldsymbol{x}|}{\|\boldsymbol{x}\|\|B\boldsymbol{x}\|} \leq \frac{\sin(2\delta)|\boldsymbol{u}^\tau\boldsymbol{\eta}|}{\frac{1}{2}\cos(\delta)\|\boldsymbol{u}\|} \leq 4\delta \leq 4\alpha.$$

□

D. Appendix B: Proofs of Theorems 4 to 6.

D.1. *Proof of Theorem 4.* Let $\boldsymbol{x}(k) =< \boldsymbol{x}, \boldsymbol{\beta}_k >$ where $\boldsymbol{\beta}_k = (0, .., 1, ...0) \in \mathbb{R}^p$ with the only 1 at the k-th position. Recall that

$$\begin{aligned}\mathcal{T} &= \left\{ k \;\middle|\; \mathbb{E}[\boldsymbol{x}(k)|y] \text{ is not constant.} \right\} \\ \mathcal{I}_p(t) &= \left\{ k \;\middle|\; var_H(\boldsymbol{x}(k)) > t \right\} \\ \mathcal{E}_p(t) &= \left\{ k \;\middle|\; var_H(\boldsymbol{x}(k)) \leq t \right\}.\end{aligned}$$

and $|\mathcal{T}| \leq Cs$ for some positive constant C. Since $var(\mathbb{E}[\boldsymbol{x}(k)|y]) > \frac{C}{s^\omega}$, we may choose $t = \frac{a}{s^\omega}$ for sufficiently small positive constant a such that for any $k \in \mathcal{T}$, $t < \frac{1}{2}var(\mathbb{E}[\boldsymbol{x}(k)|y]$. According to Lemma 1 and the Bonferroni's inequality, we have

$$\begin{aligned}\mathbb{P}\left(\mathcal{T}^c \subset \mathcal{E}_p(t)\right) &\geq 1 - \sum_{k\in\mathcal{T}^c} \mathbb{P}\left(\ var_H(\boldsymbol{x}(k)) > t\right) \\ &\geq 1 - C_1 \exp\left(-C_2\frac{ct}{H} + C_3\log(H) + \log(p-s)\right).\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}(\mathcal{T} \subset I_p(t)) &\geq \mathbb{P}\left(\bigcap_{k\in\mathcal{T}}\left\{\ var_H(\boldsymbol{x}(k)) \geq \frac{1}{2}var(\boldsymbol{m}(k,y))\ \right\}\right)\\ &\geq 1-\sum_{k\in\mathcal{T}}\mathbb{P}\left(var_H\left(\boldsymbol{x}(k)\right) < \frac{1}{2}var\left(\boldsymbol{m}(k,y)\right)\right)\\ &\geq 1-C_1\exp\left(-C_2\frac{c\ var(\boldsymbol{m}(k,y))}{H}+C_3\log(H)+\log\left(Cs\right)\right),\end{aligned}$$

i.e., we have (11) and (12) hold. □

D.2. *Proof of Theorem 5* By choosing H, c and $t = \frac{a}{s^\omega}$ properly, from Theorem 4, we have

$$P(\widehat{\mathcal{T}} = \mathcal{T}) \geq 1 - C_1\exp\left(-C_2\frac{n}{H^2 s^\omega} + C_3\log(H) + \log(p-s)\right)$$

for some positive constants $C_i, i = 1, 2, 3$. When $\widehat{\mathcal{T}} = \mathcal{T}$, we have $|\widehat{\mathcal{T}}| = O(s)$. For the n samples $(Y_i, X_i^{\widehat{\mathcal{T}}})$, apply Theorem 1, we have

$$\|e(\widehat{\boldsymbol{\Lambda}}_H^{\mathcal{T},\mathcal{T}}) - \boldsymbol{\Lambda}_p\|_2 \leq \|\widehat{\boldsymbol{\Lambda}}_H^{\mathcal{T},\mathcal{T}} - \boldsymbol{\Lambda}_p^{\mathcal{T},\mathcal{T}}\|_2 \leq O_P(\frac{1}{H^\vartheta} + \frac{H^2 s}{n} + \sqrt{\frac{H^2 s}{n}}).$$

In particular, with probability converging to one, we have

$$\|e(\widehat{\boldsymbol{\Lambda}}_H^{\mathcal{T},\mathcal{T}}) - \boldsymbol{\Lambda}_p\|_2 \to 0 \text{ as } n \to \infty.$$

D.3. *Proof of Theorem 6.* The proof is almost identical to the proof of Theorem 2, except that we additionally need to use Theorem 1 in Bickel and Levina [2008].

E. Appendix C

E.1. *Assisting Lemmas*

DEFINITION 2. *A set of random variables $x_1, ..., x_n$ can be treated as i.i.d random samples from a random variable x, if for any n variates symmetric function $f(w_1, ..., w_n)$, $f(x_1, ..., x_n)$ is identically distributed as $f(z_1, ..., z_n)$ where $z_1, ..., z_n$ are i.i.d random samples from x.*

LEMMA 9. *Let (x_i, y_i) be n i.i.d random samples from a joint distribution (x, y). Sort these samples according to the order statistics of y_i's and denote*

the sorted samples by $(x_{(1)}, y_{(1)}), (x_{(2)}, y_{(2)}), ..., (x_{(n)}, y_{(n)})$. *Then for any* a, b $(1 \le a \le b+1 \le n)$, $x_{(a+1)}, ..., x_{(b)}$ *can be treated as* $b-a$ *i.i.d samples from* $x \Big| \left(y \in [y_{(a)}, y_{(b+1)}]\right)$.

PROOF. In fact, we only need to prove that $y_{(a+1)}, \cdots, y_{(b)}$ can be treated as $b-a$ i.i.d. samples of $y \Big| (y \in [y_{(a)}, y_{(b+1)}])$. The latter only needs to be proved for uniform distribution which can be verified directly.

COROLLARY 1. *In the slicing inverse regression contexts, recall that* S_h *denotes the* h*-th interval* $(y_{h-1,c}, y_{h,c}]$ *for* $2 \le h \le H-1$ *and* $S_1 = (-\infty, y_{1,c}]$, $S_H = (y_{H-1,c}, \infty)$. *We have that* $x_{h,i}, i = 1, \cdots, c-1$ *can be treated as* $c-1$ *random samples of* $x \Big| (y \in S_h)$ *for* $h = 1, ..., H-1$ *and* $x_{H,1}, ..., x_H$ *can be treated as* c *random samples of* $x \Big| (y \in S_H)$.

LEMMA 10. *Suppose that* (x, y) *are defined over* σ*-finite space* $\mathcal{X} \times \mathcal{Y}$ *and* g *is a non-negative function such that* $\mathbb{E}[g(x)]$ *exists. For any fixed positive constants* $C_1 < 1 < C_2$, *there exists a constant* C *which only depends on* C_1, C_2 *such that for any partition* $\mathbb{R} = \bigcup_{h=1}^{H} S_h$ *where* S_h *are intervals satisfying*

$$\frac{C_1}{H} \le \mathbb{P}(y \in S_h) \le \frac{C_2}{H}, \forall h, \tag{E.1}$$

we have

$$\sup_h \mathbb{E}(g(x) \Big| y \in S_h') \le CH\mathbb{E}[g(x)].$$

PROOF. According to Fubini's Theorem, for any h,

$$\begin{aligned}\mathbb{E}[g(x)] =& \sum_k \mathbb{P}(y \in S_k) \int_{\mathcal{X}} g(x) p(x|y \in S_k) dx \\ \ge& \mathbb{P}(y \in S_h) \int_{\mathcal{X}} g(x) p(x|y \in S_h) dx.\end{aligned}$$

Due to the condition (E.1), there exists a positive constant C such that

$$\int_{\mathcal{X}} g(x) p(x|y \in S_h) dx \le CH\mathbb{E}[g(x)].$$

□

COROLLARY 2. *Let $\boldsymbol{x}$ be a multivariate random variable with covariance matrix $\boldsymbol{\Sigma}$. For any partition satisfying (E.1), there exists a constant C such that*

$$var(\boldsymbol{\beta}^\tau \boldsymbol{x}|_{y\in S'_h}) \leq CHvar(\boldsymbol{\beta}^\tau \boldsymbol{x}), \text{for any unit vector } \boldsymbol{\beta},$$

and

$$\lambda_{max}\left(var\left(\boldsymbol{x}\middle| y \in S'_h\right)\right) \leq CH\lambda_{max}\left(var\left(\boldsymbol{x}\right)\right).$$

COROLLARY 3. *Let x be a sub-Gaussian random variable which is upper-exponentially bounded by K. Then for any partition satisfying (E.1), there exists a constant C such that*

$$\mathbb{E}[\exp\left(\frac{x^2}{K^2}\right)\Big| y \in S'_h] \leq CH\mathbb{E}[\exp\left(\frac{x^2}{K^2}\right)].$$

Recall the definition of the random intervals $S_h, h = 1, 2, \cdots, H$ and random variable $\delta_h = \delta_h(\omega) = \int_{y\in S_h(\omega)} f(y)dy$.

LEMMA 11. *Define the event $E(\epsilon) = \left\{ \omega \;\middle|\; |\delta_h - \frac{1}{H}| > \epsilon, \forall h \right\}$. There exists a positive constant C such that, for any $\epsilon > \frac{4}{Hc-1}$ we have*

$$P(E(\epsilon)) \leq CH^2\sqrt{Hc+1}\exp(-(Hc+1)\frac{\epsilon^2}{32}) \tag{E.2}$$

for sufficient large H and c.

PROOF. The proof is deferred to the end of this paper.

E.2. *Some Results from Random Matrices Theory.* We collect some direct corollaries of the non-asymptotic random matrices theory (e.g., Rudelson and Vershynin [2013]).

LEMMA 12. *Let $\boldsymbol{M}$ be any $p \times n$ matrix ($n > p$) whose columns $\boldsymbol{M}_i$ are independent sub-Gaussian random vectors in $\mathbb{R}^p$ with second moment $\mathbf{I}_p$ and $\lambda^+_{sing,min}(\boldsymbol{M})$, $\lambda_{sing,max}(\boldsymbol{M})$ be the minimal non-zero and maximal singular value of $\boldsymbol{M}$. Then for every t, with probability at least $1 - 2\exp(-C't^2)$, we have :*

$$\sqrt{n} - C\sqrt{p} - t \leq \lambda^+_{sing,min}(\boldsymbol{M}) \leq \lambda_{sing,max}(\boldsymbol{M}) \leq \sqrt{n} + C\sqrt{p} + t.$$

LEMMA 13. *Let $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ be n i.i.d. samples from a p-dimensional sub-Gaussian random variable with covariance matrix $\boldsymbol{\Sigma}$ and $\rho = \frac{p}{n}$. If there exists positive constants C_1 and C_2 such that*

$$C_1 \leq \lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{x}}) \leq \lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{x}}) \leq C_2.$$

Let $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} = \frac{1}{n}\sum_i \boldsymbol{x}_i \boldsymbol{x}_i^\tau$. Then

$$\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}}\|_2 \to 0 \text{ if } \rho = 0 \text{ when } n \to \infty.$$

It is also easy to see that, given the boundedness condition on $\boldsymbol{\Sigma}_X$, $\|\widehat{\boldsymbol{\Sigma}}_X^{-1} - \boldsymbol{\Sigma}_X^{-1}\|_2 \to 0$ if $\rho = \frac{p}{n} \to 0$ when $n \to \infty$.

PROOF. Let $\boldsymbol{x}_i = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{1/2}\boldsymbol{m}_i$ where $\boldsymbol{m}_i$ is sub-Gaussian random variable with covariance matrix $\mathbf{I}_p$ and $\boldsymbol{M} = (\boldsymbol{m}_1, ..., \boldsymbol{m}_n)$. From Lemma 12, we have

$$\|\frac{1}{n}\boldsymbol{M}\boldsymbol{M}^\tau - \mathbf{I}_p\|_2 \to 0$$

and

$$\|\widehat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\|_2 = \|\boldsymbol{\Sigma}^{-1/2}\|_2 \|\frac{1}{n}\boldsymbol{M}\boldsymbol{M}^\tau - \mathbf{I}_p\|_2 \|\boldsymbol{\Sigma}^{-1/2}\|_2 \to 0,$$

with probability converges to 1 as $n \to \infty$ because

$$\lambda_{max}\left(\frac{1}{n}\boldsymbol{M}\boldsymbol{M}^\tau\right) \leq \left(1 + \frac{(C+1)\sqrt{p}}{\sqrt{n}}\right)^2$$

and

$$\lambda_{min}\left(\frac{1}{n}\boldsymbol{M}\boldsymbol{M}^\tau\right) \geq \left(1 - \frac{(C+1)\sqrt{p}}{\sqrt{n}}\right)^2$$

with probability at least $1 - 2\exp(-C'p)$. □

The following lemma is well known in the non-asymptotic random matrix theory (Vershynin [2010] Proposition 5.34) which is slightly different from the Lemma 12.

LEMMA 14. *Let $\boldsymbol{E}_{p\times H}$ be a $p \times H$ matrix, whose entries are independent standard normal random variables. Then for every $t \geq 0$, with probability at least $1 - 2\exp(-t^2/2)$, we have :*

$$\lambda^+_{sing,min}(\boldsymbol{E}_{p\times H}) \geq \sqrt{p} - \sqrt{H} - t,$$

and

$$\lambda_{sing,max}(\boldsymbol{E}_{p\times H}) \leq \sqrt{p} + \sqrt{H} + t.$$

COROLLARY 4. *We have*

$$\frac{1}{2}\left(\sqrt{p}-\sqrt{H}\right) \leq \lambda^{-}_{sing,min}\left(\boldsymbol{E}_{p\times H}\right) \leq \lambda_{sing,max}\left(\boldsymbol{E}_{p\times H}\right) \leq \frac{3}{2}\left(\sqrt{p}+\sqrt{H}\right).$$

with probability converging to one, as $n \to \infty$.

PROOF. Choosing $t = \sqrt{p}/2$, according to Lemma 14, we have:

$$\mathbb{P}\left(\frac{\lambda_{max}(E_H)}{\sqrt{p}+\sqrt{H}} \leq \frac{3}{2}\right) \geq \mathbb{P}\left(\frac{\lambda_{max}(E_H)}{\sqrt{p}+\sqrt{H}} \leq 1+\frac{\sqrt{p}}{2\sqrt{p}+2\sqrt{H}}\right)$$

and

$$\mathbb{P}\left(\frac{\lambda^{+}_{min}(E_H)}{\sqrt{p}-\sqrt{H}} \geq \frac{1}{2}\right) \geq \mathbb{P}\left(\frac{\lambda_{max}(E_H)}{\sqrt{p}-\sqrt{H}} \geq 1-\frac{\sqrt{p}}{2\sqrt{p}-2\sqrt{H}}\right)$$

with probability at least $1-2\exp(-p/8)$. i.e., With probability converging to one, we have

$$\frac{1}{2}(\sqrt{p}-\sqrt{H}) \leq \lambda^{-}_{min}(E_{p\times H}) \leq \lambda_{max}(E_{p\times H}) \leq \frac{3}{2}(\sqrt{p}+\sqrt{H}).$$

□

E.3. *Basic Properties of sub-Gaussian random variables.* We rephrased several equivalent definitions of the sub-Gaussian distribution here (See e.g., Vershynin [2010]):

DEFINITION 3. *Let* x *be a random variable. Then the following properties are equivalent with parameters* K_i*'s differing from each other by at most an absolute constant factor,*

1. *Tails:* $\mathbb{P}(|x| > t) \leq \exp(1-t^2/K_1^2)$ *for all* $t \geq 0$.
2. *Moments:* $(\mathbb{E}[|x|^p])^{1/p} \leq K_2\sqrt{p}$ *for all* $p \geq 1$.
3. *Super-exponential moment:* $\mathbb{E}\exp(x^2/K_3^2) \leq e$.

Moreover, if $\mathbb{E}[x] = 0$, *then the properties* $1-3$ *are also equivalent to the following one:*

4. *Moment generating function:* $\mathbb{E}[\exp(tx)] \leq \exp(t^2K_4^2)$.

DEFINITION 4. *For a sub-Gaussian random variable* x *with the constants* $K_i, i = 1,2,3,4$ *given in Definition 3, we will call a constant* K *an upper-exponential bound of* x *or* x *is upper-exponentially bounded by* K *if* $K > \max_i\{K_1, K_2, K_3, K_4\}$.

We summarize some properties regarding the sub-Gaussian distributions into the following lemmas.

LEMMA 15. *Let $\delta_1, ..., \delta_n$ be n (not necessarily independent or with mean zero) sub-Gaussian random variables upper-exponentially bounded by K.*

i) *$\frac{1}{n}\sum_{i=1}^n \delta_i$ is sub-Gaussian and upper-exponentially bounded by K.*

ii) *$\delta_1 - \mathbb{E}[\delta_1]$ is sub-Gaussian upper-exponentially bounded by $2K$.*

iii) *If they are independent and with mean zero , then $\frac{1}{\sqrt{n}}\sum_i \delta_i$ is sub-Gaussian and upper-exponentially bounded by K.*

iv) *If they are i.i.d., then we have the concentration inequality:*

$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^n \delta_i}{n} - \mathbb{E}[x]\right| > t\right) \leq 2\exp\left(\frac{-nt^2}{2K^2e + 2tK}\right).$$

PROOF. **i)** follows from the linear property of expectation and the the convexity of exponential function. i.e.,

$$\mathbb{E}[\exp(\frac{1}{nK^2}\sum_i \delta_i^2)] \leq \mathbb{E}[\frac{1}{n}\sum_h \exp(\frac{\delta_i^2}{K^2})] \leq \max_i \mathbb{E}[\exp(\frac{\delta_i^2}{K^2})] \leq e.$$

ii) From Definition 3, we know that $|\mathbb{E}[\delta_i]| \leq K$ which gives us the desired upper-exponential bound of $\delta_i - \mathbb{E}[\delta_i]$.
iii) is trivial as $\delta_1, \cdots, \delta_c$ are independent and with mean zero.
iv) Since δ_1 is sub-Gaussian upper-exponentially bounded by K, we have:

$$\begin{aligned}
\mathbb{E}[|\delta_1|^p] &= \int_0^\infty pt^{p-1}\mathbb{P}(|\delta_1| > t)dt \leq \int_0^\infty pt^{p-1}\exp\left(1 - \frac{t^2}{K^2}\right)dt \\
&= \frac{ep}{2}\Gamma(\frac{p}{2})K^p && \text{for any } p \geq 1 \\
&\leq p!K^{p-2}\frac{(K^2e)}{2} && \text{for any } p \geq 2
\end{aligned}$$

Recall the well known Bernstein inequality.

LEMMA 16. **(Bernstein Inequality).** *If there exists positive constants V and b such that for any integers $m \geq 2$,*

$$E[|\delta_1|^m] \leq m!b^{m-2}V/2$$

then

$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^n \delta_i}{n} - \mathbb{E}[x]\right| > t\right) \leq 2\exp\left(\frac{-nt^2}{2V + 2tb}\right). \tag{E.3}$$

By chooing $V = K^2 e$ and $b = K$, we get the desired concentration inequality. □

LEMMA 17. *Suppose that (x, y) are defined over σ-finite space $\mathcal{X} \times \mathcal{Y}$ and x is sub-Gaussian with mean 0 and upper exponentially bounded by K, let $m(y) = \mathbb{E}[x|y]$, $\epsilon(y) = x - m(y)$, then we have*

i) *$m(y)$ and $\epsilon(y)$ are sub-Gaussian and upper-exponentially bounded by K and $2K$ respectively.*

ii) *Let $\mathcal{Z}$ consists of points y such that $x|_y$ is not sub-Gaussian, i.e.,*

$$\mathcal{Z} \triangleq \left\{ y \mid \exists t \in (0, t_0] \text{ such that } \int_{\mathcal{X}} \exp(tx^2) p(x|y) p(y) dx = \infty \right\},$$

then $\mathbb{P}(y \in \mathcal{Z}) = 0$.

iii) *For any fixed positive constants $C_1 < 1 < C_2$ and any partition $\mathbb{R} = \bigcup_{h=1}^{H} S_h$ where S_h are intervals satisfying*

$$\frac{C_1}{H} \leq \mathbb{P}(y \in S_h) \leq \frac{C_2}{H}, \forall h,$$

there exists a constant C such that

$$\sup_h \mathbb{P}(x|_{y \in S_h} > t) \leq CH \exp\left(1 - \frac{t^2}{K^2}\right).$$

As a direct corollary, we know that there exists a positive constant C such that

$$\mathbb{E}\left[\exp\left(\frac{(x|_{y \in S_h})^2}{2K^2}\right)\right] \leq CH,$$

and

$$\mathbb{E}\left[\left|(x|_{y \in S_h})\right|^m\right] \leq CHmK^m \Gamma(\frac{m}{2})/2.$$

vi) *Suppose that $x|_{y \in S_h}$ is defined as in* **iii**). *Let x_i, i=1,...,c be c samples from $x|_{y \in S_h}$, $\bar{x}_h = \frac{1}{c} \sum_i x_i$ and $\mu_h = \mathbb{E}[x|_{y \in S_h}]$, we have*

$$\mathbb{P}[|\bar{x}_h - \mu_h| > t] \leq 2 \exp\left(\frac{-ct^2}{2CHK^2 + 2tK}\right).$$

PROOF. **i)** By Jensen's inequality, we have

$$\mathbb{E}[\exp(t\mathbb{E}[x|y])] \leq \mathbb{E}[\mathbb{E}[\exp(tx)|y]] = \mathbb{E}[\exp(tx)] \leq \exp(t^2 K_1^2).$$

i.e., $m(y)$ is sub-Gaussian and upper-exponentially bounded by K_1. Since x , $m(y)$ is sub-Gaussian and upper-exponentially bounded by K_1, we know that $\epsilon = x - m(y)$ is sub-Gaussian and upper-exponentially bounded by $2K_1$.

ii) Let $p(x, y)$ be the joint density function of (x, y) and $p(x)$, $p(y)$ be the marginal distribution of x, y. Since x is sub-Gaussian, we know there exists $t_0 > 0$ such that

$$\int_{\mathcal{X}} \exp(tx^2) \int_{\mathcal{Y}} p(x|y)p(y)dydx \leq e \text{ for } 0 \leq t \leq t_0.$$

By Fubini Theorem, we know

$$\int_{\mathcal{Y}} p(y) \int_{\mathcal{X}} \exp(tx^2)p(x|y)dxdy \leq e \text{ for } 0 \leq t \leq t_0. \tag{E.4}$$

Recall that we have $\mathcal{Z} \triangleq \{y|\exists t \in (0, t_0] \text{ such that } \int_{\mathcal{X}} \exp(tx^2)p(x|y)p(y)dx = \infty\}$, from equation (E.4), we know $\mathbb{P}(y \in \mathcal{Z})$=0. In particular, we know that for any $y \notin \mathcal{Z}$, $x|_y$ is sub-Gaussian. However, the norm (e.g., sub-exponential norm)of $x|_y$ might be varying along with y and , as a function of y , it might be not bounded.

iii) From Lemma 10, we know that there exists a positive constant C such that

$$\int_{\mathcal{X}} \exp(tx^2)p(x|y \in S_h)dx \leq CHe.$$

For simplicity if notation, we will denote $x|_{y \in S_h}$ by z through out this lemma. So for $0 \leq t \leq t_0 = \frac{1}{K}$, we have

$$\mathbb{P}(z > a) \leq \frac{\mathbb{E}[\exp(tz)^2)]}{\exp(t^2a^2)} \leq CHe \exp\left(-t^2a^2\right).$$

From the above tail bounds, we have that for any integer $m > 0$

$$\begin{aligned}\mathbb{E}[|z|^m] &= \int_0^\infty \mathbb{P}(|z| > t)(m)t^{m-1}dt \leq CHm \int_0^\infty exp(-\frac{t^2}{K^2})t^{m-1}dt \\ &\leq CHmK^m(m/2)/2.\end{aligned}$$

We then have

$$\begin{aligned}\mathbb{E}\left[\exp\left(tz^2\right)\right] &\leq \sum_{m=0}^{\infty} \frac{\mathbb{E}\left[t^m z^{2m}\right]}{m!} \leq \sum_{m=0}^{\infty} \frac{\mathbb{E}\left[t^m z^{2m}\right]}{m!} \\ &\leq \sum_{m=0}^{\infty} \frac{t^m CHmK^{2m}\Gamma(m)}{m!} = CH \sum_{m=0}^{\infty} t^m K^{2m}\end{aligned}$$

From which we know if $0 \leq t < \frac{1}{2}K^{-2}$, the R.H.S is bounded by CH for a positive constant C.

vi) From the previous proof, we know that for any integer $m \geq 2$

$$\mathbb{E}[|z|^m] \leq CHm!K^m = m!K^{m-2}(2CHK^2)/2.$$

By the Bernstein inequality (E.3), we have:

$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^{c} z_i}{c} - \mathbb{E}[z]\right| > t\right) \leq 2\exp\left(\frac{-ct^2}{2CHK^2 + 2tK}\right).$$

□

LEMMA 18. *Let $z_i, i = 1, \cdots, n$ be i.i.d. samples of a sub-Gaussian distribution exponentially upper bounded by K, then there exist positive constants C_1, C_2 such that, if $\sqrt{n}\epsilon \to \infty$, we have*

$$\mathbb{P}(|\frac{1}{n}\sum_i (z_i - \bar{z})^2 - var(z)| > \epsilon) \leq C_1 \exp(-C_2 \frac{\epsilon\sqrt{n}}{K^2}),$$

where $\bar{z} = \frac{1}{n}\sum_i z_i$.

PROOF. Recall the following Hanson-Wright inequality in Rudelson and Vershynin [2013]

LEMMA 19. *Let $\boldsymbol{v} = (\boldsymbol{x}(1), \cdots, \boldsymbol{x}(n))$ be a sub-Gaussian random vector with independent components $\boldsymbol{x}(\boldsymbol{\beta})$ such that $\mathbb{E}[\boldsymbol{x}(\boldsymbol{\beta})] = 0$ and $\|\boldsymbol{x}(\boldsymbol{\beta})\|_{\psi_2} \leq K$. Let $\boldsymbol{A}$ be an $n \times n$ matrix. Then there exists a positive constant C such that for any $t > 0$,*

$$\mathbb{P}\{|\boldsymbol{x}^\tau \boldsymbol{A}\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}^\tau A\boldsymbol{x}]| > t\} \leq 2\exp\left(-C \ \min\left(\frac{t^2}{K^4\|\boldsymbol{A}\|_{HS}^2}, \frac{t}{K^2\|\boldsymbol{A}\|_{HS}}\right)\right).$$

Here the ψ_2 norm of a random variable z is defined as $\|z\|_{\psi_2} \triangleq \sup_p p^{-1/2}(\mathbb{E}|z|^p)^{1/p}$ and the HS norm of a matrix $\boldsymbol{A}$ is defined as $\|\boldsymbol{A}\|_{HS} = (\sum_{i,j}|a_{i,j}|^2)^{1/2}$.

Since

$$\begin{aligned}
&\mathbb{P}\left(|\frac{1}{n}\sum_i (z_i - \bar{z})^2 - var(z)| > 2\epsilon\right) \\
&= \mathbb{P}\left(|\frac{1}{n}\sum_i (z_i - \mathbb{E}[z])^2 - (\mathbb{E}[z] - \bar{z})^2 - \mathbb{E}[(z - E[z])^2]| > 2\epsilon\right) \\
&\leq \mathbb{P}\left(|\frac{1}{n}\sum_i (z_i - \mathbb{E}[z])^2 - \mathbb{E}[(z - \mathbb{E}[z])^2]| > \epsilon\right) + \mathbb{P}\left((\mathbb{E}[z] - \bar{z})^2 > \epsilon\right),
\end{aligned}$$

and $z_i - \mathbb{E}[z]$ are sub-Gaussian with mean 0, from Lemma 19 by choosing $\boldsymbol{A} = \frac{1}{n}\mathbf{I}_p$ and $\boldsymbol{z}^\tau = (z_1 - \mathbb{E}[z], .., z_p - \mathbb{E}[z])$, we have

$$\mathbb{P}\left(|\frac{1}{n}\sum_i (z_i - \mathbb{E}[z])^2 - \mathbb{E}[(z - \mathbb{E}[z])^2]| > \epsilon\right) \leq 2\exp\left(-C\frac{\sqrt{n}\epsilon}{K^2}\right), \tag{E.5}$$

since $\sqrt{n}\epsilon \to \infty$.

The following follows from the usual deviation argument:

$$\mathbb{P}\left((\mathbb{E}[z] - \bar{z})^2 > \epsilon\right) \leq C_1 \exp\left(-C_2 n\epsilon\right).$$

Combining with the estimate (E.5), we know there exists positive constants C_1 and C_2 such that

$$\mathbb{P}\left(|\frac{1}{n}\sum_i (z_i - \bar{z})^2 - var(z)| > \epsilon\right) \leq C_1 \exp\left(-C_2\frac{\sqrt{n}\epsilon}{K^2}\right),$$

for sufficiently large n since $\sqrt{n}\epsilon \to \infty$. □

E.4. *Proof of Lemma 11.* We only need to prove this lemma for n i.i.d. sample y_i's from a uniform distribution over $[0, 1]$. We slightly change the notation of order statistics $y_{(i)}$ to $y_{(i,n)}$ so that we can keep track of the sample size. Since y is uniform distribution on $[0, 1]$, it is well known that $y_{(i,n)} \sim Beta(i, n - i + 1)$ with expectation $\frac{i}{n+1}$ and mode $\frac{i-1}{n-1}$. Lemma 11 is a direct corollary of the following lemma.

LEMMA 20. *Suppose there are $n = Hc$ i.i.d. samples from uniform distribution over $[0, 1]$, when H,c are sufficiently large, we have the following large deviation inequalities of $y_{(kc,Hc)}$, $k = 1, \cdots, (H - 1)$.*

i) There exists a positive constant C, such that for any $\epsilon > \frac{1}{Hc-1}$, we have

$$\mathbb{P}\left(y_{(kc,Hc)} > \frac{k}{H} + \epsilon\right) \leq CH\sqrt{Hc+1}\exp\left(-(Hc+1)\frac{\epsilon^2}{2}\right);$$

ii) There exists a positive constant C, such that for any $\epsilon > \frac{2}{Hc-1}$, we have

$$\mathbb{P}\left(y_{(kc,Hc)} < \frac{k}{H} - \epsilon\right) \leq CH\sqrt{Hc+1}\exp\left(-(Hc+1)\frac{\epsilon^2}{8}\right);$$

iii) *Let* $\delta(k,H,c) = |y_{((k-1)c,Hc)} - y_{(kc,Hc)}|$ *, for* $2 \le k \le H-1$, $\delta(1,H,c) = |y_{(c,Hc)}|$ *and* $\delta(H,H,c) = |1 - y_{((H-1)c,Hc)}|$. *There exists a positive constant* C, *such that for any* $\epsilon > \frac{4}{Hc-1}$ *, we have for any* $1 \le k \le H$*:*

$$\mathbb{P}\left(|\delta(k,H,c) - \frac{1}{H}| > \epsilon\right) \le CH\sqrt{Hc+1}\exp\left(-(Hc+1)\frac{\epsilon^2}{32}\right).$$

We will prove Lemma 20 later. Assuming it, we have

$$\begin{aligned}\mathbb{P}\left(E(\epsilon)\right) &\le \sum_{k=1}^{H}\mathbb{P}\Big(|\delta(k,H,c) - \frac{1}{H}| > \epsilon\Big) \\ &\le CH^2\sqrt{Hc+1}\exp\left(-(Hc+1)\frac{\epsilon^2}{32}\right).\end{aligned}$$

□

E.4.1. *Proof of Lemma 20. The first part:* For any $1 \le k \le H-1$, we note that

$$\begin{aligned}&\mathbb{P}\left(y_{(kc,Hc)} > \frac{k}{H} + \epsilon\right) \le \mathbb{P}\left(y_{(kc,Hc)} > \frac{kc}{Hc+1} + \epsilon\right) \\ &= \frac{1}{B(kc, Hc-kc+1)}\int_{x > \frac{kc}{Hc+1}+\epsilon} x^{kc-1}(1-x)^{Hc-kc}dx.\end{aligned}$$

When $\epsilon > \frac{1}{Hc-1}$, we know the mode $x_M = \frac{kc-1}{Hc-1} < x_D \triangleq \frac{kc}{Hc+1} + \epsilon$, so we have

$$\begin{aligned}\mathbb{P}\left(y_{(kc,Hc)} > \frac{k}{H} + \epsilon\right) &\le \frac{(x_D)^{kc-1}(1-x_D)^{Hc-kc+1}}{B(kc, Hc-kc+1)} \\ &\le H\frac{(x_D)^{kc}(1-x_D)^{Hc-kc+1}}{B(kc, Hc-kc+1)}.\end{aligned} \tag{E.6}$$

The last inequality due to $Hx_D \ge 1$. If $\epsilon + \frac{k}{H} \ge 1$, then $\mathbb{P}(y_{(kc,Hc)} > \frac{k}{H} + \epsilon) = 0$ and Lemma 20 holds automatically.

We may assume that $\epsilon + \frac{k}{H} < 1$ below. Let us denote the right hand side of (E.6) by A, then

$$\begin{aligned}\log(A) =& \log(H) + kc\log(E+\epsilon) + (Hc-kc+1)\log(1-E-\epsilon) \\ &+ \log(Hc+1)! - \log(kc)! - \log(Hc-kc+1)! \\ &- \log(Hc+1) + \log(kc) + \log(Hc-kc+1),\end{aligned}$$

where $E = \frac{kc}{Hc+1}$. According to the Stirling formula:

$$\log(n!) = n\log(n) - n + \frac{1}{2}\log(2\pi n) + O(\frac{1}{n}),$$

when m is sufficiently large we have:

$$\begin{aligned}\log(A) =& \log(H) + (Hc+1)(E\log(1+\frac{\epsilon}{E}) + (1-E)\log(1-\frac{\epsilon}{1-E})) \\ &- \frac{1}{2}(\log(Hc+1) - \log(kc) - \log(Hc-kc+1)) \\ &- \frac{1}{2}\log(2\pi) + O(\frac{1}{kc}) + O(\frac{1}{Hc-kc+1}) \\ \leq& \log(H) - \frac{(Hc+1)\epsilon^2}{2(1-E)} - \frac{1}{2}\log(2\pi) \\ &- \frac{1}{2}(\log(Hc+1) - \log(kc) - \log(Hc-kc+1)) + O(\frac{1}{c}) \\ \leq& \log(H) - \frac{(Hc+1)\epsilon^2}{2(1-E)} - \frac{1}{2}(\log(Hc+1) - \log(kc) - \log(Hc-kc+1)),\end{aligned} \tag{E.7}$$

where we use the fact that $\frac{kc}{Hc+1} \leq E + \epsilon < 1$ and the following elementary lemma, which can be proved by the Taylor expansion:

LEMMA 21. *Suppose a, b are positive numbers such that $a + b = 1$, then for any $0 < \epsilon < b$, we have:*

$$a\log(1+\frac{\epsilon}{a}) + b\log(1-\frac{\epsilon}{b}) \leq -\frac{\epsilon^2}{2b}.$$

Now we know that there exists a positive constant C such that for any $1 \leq k \leq H-1$ and for any $\epsilon > \frac{1}{Hc-1}$, the following holds:

$$\begin{aligned}&\mathbb{P}\Big(y_{(kc,Hc)} > \frac{k}{H} + \epsilon\Big) \\ &\qquad \leq CH\sqrt{\frac{(kc)(Hc-kc+1)}{Hc+1}}\exp\left(-(Hc+1)\frac{\epsilon^2}{2(1-E)}\right) \\ &\qquad \leq CH\sqrt{Hc+1}\exp\left(-(Hc+1)\frac{\epsilon^2}{2(1-E)}\right) \\ &\qquad \leq CH\sqrt{Hc+1}\exp\left(-(Hc+1)\frac{\epsilon^2}{2}\right).\end{aligned}$$

The last inequality follows from $\frac{\epsilon^2}{1-E} \geq \epsilon^2$ since $\frac{1}{H+1} \leq E \leq \frac{H-1}{H}$.

The second part: The proof of the second part is similar. For completeness, we sketch some calculations below. For any $1 \le k \le H-1$, when $\epsilon > \frac{2}{Hc-1}$, we have

$$\begin{aligned}&\mathbb{P}\left(y_{(kc,Hc)} < \frac{k}{H} - \epsilon\right) \le \mathbb{P}\left(y_{(kc,Hc)} < \frac{kc}{Hc+1} - \epsilon/2\right)\\ &= \frac{1}{B(kc, Hc-kc+1)} \int_{x < \frac{kc}{Hc+1} - \epsilon} x^{kc-1}(1-x)^{Hc-kc} dx.\end{aligned}$$

Since $\epsilon > \frac{1}{Hc-1}$, we know the mode $x_M = \frac{kc-1}{Hc-1} > x_{D'} \triangleq \frac{kc}{Hc+1} - \epsilon/2$, so we have

$$\begin{aligned}\mathbb{P}\left(y_{(kc,Hc)} \le \frac{k}{H} - \epsilon\right) \le& \frac{(x_{D'})^{kc}(1-x_{D'})^{Hc-kc}}{B(kc, Hc-kc+1)}\\ \le& H\frac{(x_{D'})^{kc}(1-x_{D'})^{Hc-kc+1}}{B(kc, Hc-kc+1)}.\end{aligned}$$

The last inequality due to $H(1-x_{D'}) \ge 1$.

The rest is similar to the first part. We have that for any $1 \le k \le H-1$ and for any $\epsilon > \frac{2}{Hc-1}$,

$$\mathbb{P}\left(y_{(kc,Hc)} < \frac{k}{H+1} - \epsilon\right) \le CH\sqrt{Hc+1}\exp\left(-(Hc+1)\frac{\epsilon^2}{8}\right). \tag{E.8}$$

The third part: The third part is a direct corollary of the first two parts. Note that for any $2 \le k \le H-2$, for any $\epsilon > \frac{4}{Hc-1}$

$$\begin{aligned}&\mathbb{P}\Big(|\delta(k,H,c) - \frac{1}{H}| > \epsilon\Big)\\ &\quad= \mathbb{P}\left(\left|y_{(k+1)c,Hc} - \frac{k+1}{H} - (y_{kc,Hc} - \frac{k}{H})\right| > \epsilon\right)\\ &\quad\le \mathbb{P}\left(|y_{(k+1)c,Hc} - \frac{k+1}{H}| > \frac{\epsilon}{2}\right) + \mathbb{P}\left(|y_{kc,Hc} - \frac{k}{H}| > \frac{\epsilon}{2}\right)\\ &\quad\le CH\sqrt{Hc+1}\exp\left(-(Hc+1)\frac{\epsilon^2}{32}\right).\end{aligned}$$

When $k = 1$, we have

$$\begin{aligned}\mathbb{P}\Big(|\delta(1,H,c) - \frac{1}{H}| > \epsilon\Big) &= \mathbb{P}\left(|y_{c,Hc} - \frac{1}{H}| > \epsilon\right) \\ &\leq \mathbb{P}\left(y_{c,Hc} - \frac{1}{H} > \epsilon\right) + \mathbb{P}\left(y_{c,Hc} - \frac{1}{H} < -\epsilon\right) \\ &\leq CH\sqrt{Hc+1}\exp\left(-(Hc+1)\frac{\epsilon^2}{8}\right) \\ &\leq CH\sqrt{Hc+1}\exp\left(-(Hc+1)\frac{\epsilon^2}{32}\right).\end{aligned}$$

When $k = H$, we have

$$\begin{aligned}\mathbb{P}\Big(|\delta(H-1,H,c) - \frac{1}{H}| > \epsilon\Big) &= \mathbb{P}\left(|y_{(H-1)c,Hc} - \frac{H-1}{H}| > \epsilon\right) \\ &\leq \mathbb{P}\left(y_{(k+1)c,Hc} - \frac{H-1}{H} > \epsilon\right) + \mathbb{P}\left(y_{(H-1)c,Hc} - \frac{H-1}{H} < -\epsilon\right) \\ &\leq CH\sqrt{Hc+1}\exp\left(-(Hc+1)\frac{\epsilon^2}{8}\right) \\ &\leq CH\sqrt{Hc+1}\exp\left(-(Hc+1)\frac{\epsilon^2}{32}\right).\end{aligned}$$

□

References

P. J. Bickel and E. Levina. Covariance regularization by thresholding. *The Annals of Statistics*, pages 2577–2604, 2008.

I. M. Johnstone and A. Y. Lu. Sparse principal components analysis. 2004.

M. Rudelson and R. Vershynin. Hanson-wright inequality and sub-gaussian concentration. 2013.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.